UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS HÍDRICOS

IGOR SANTOS SILVA

INTELIGÊNCIA ARTIFICIAL PARA AVALIAÇÃO DA QUALIDADE DA ÁGUA

São Cristóvão, SE.

2019

IGOR SANTOS SILVA

INTELIGÊNCIA ARTIFICIAL PARA AVALIAÇÃO DA QUALIDADE DA ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Recursos Hídricos como um dos requisitos de obtenção do título de Mestre em Recursos Hídricos.

Orientador: Prof. Dr. Carlos Alexandre Borges Garcia

Coorientadora: Profa. Dra. Helenice Leite Garcia

São Cristóvão, SE

2019

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Silva, Igor Santos
S586i Inteligência artificial para avaliação da qualidade da água / Igor Santos Silva ; orientador Carlos Alexandre Borges Garcia. – São Cristóvão, SE, 2019.
104 f. : il.

Dissertação (mestrado em Recursos Hídricos) – Universidade Federal de Sergipe, 2019.

1. Recursos hídricos – Administração. 2. Água – Qualidade – Medição. 3. Clorofila. 4. Reservatórios – Sergipe. 5. Aprendizado do computador. I. Garcia, Carlos Alexandre Borges, orient. II. Título.

CDU 556.18:004.85(813.7)

IGOR SANTOS SILVA

INTELIGÊNCIA ARTIFICIAL PARA AVALIAÇÃO DA QUALIDADE DA ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Recursos Hídricos como um dos requisitos de obtenção do título de Mestre em Recursos Hídricos.

Aprovada: 22 de fevereiro de 2019.

Prof. Dr. Carlos Alexandre Borges Garcia
(Orientador)

Prof. Dr. José do Patrocínio Hora Alves
Programa de Pós-Graduação em Recursos Hídricos

Prof. Dr. Jefferson Arlen Freitas
Departamento de Engenharia Ambiental

São Cristóvão, SE

2019

Deus é o nosso refúgio e fortaleza, socorro bem presente na angústia. (Sm 46:1)

Dedico este trabalho aos meus pais, Ocsicnarf e José, por serem minha fonte de inspiração.

AGRADECIMENTOS

Em primeiro lugar a Deus, por abençoar cada passo meu e por toda Sua grandiosidade em minha vida, ensinando-me a todo momento a ter paciência e confiar nEle. Sem Ele nada seria possível.

Aos meus pais, Miu e Zé, e minha irmã, Illa, por acreditarem em mim em todo e a todo momento, principalmente, nos momentos em que nem eu mesmo acreditava. Obrigado pelas palavras de força, pela presença e pela paciência durante essa jornada.

A Prof. Dr. Carlos Alexandre Borges Garcia pelo companheirismo, "deboísmo", por acreditar em mim e sempre passando a certeza de que tudo vai dar certo.

A Profa. Dra. Helenice Leite Garcia por acreditar no trabalho, pela dedicação total, pela confiança, pela paciência comigo, pelos ensinamentos, pela imensa amizade e por proporcionar momentos de aprendizados não somente acadêmicos.

Ao Prof. Dr. Inajá Francisco de Sousa por toda sua dedicação e integridade em prol do programa e dos alunos do mestrado visando a melhora do mesmo.

Ao Prof. Dr. José do Patrocínio Hora Alves pelo auxílio na aquisição de dados, e principalmente, com seus ensinamentos e apoio. Sinto-me honrado por sua colaboração.

Aos amigos Paulo, Keilla e Mariana por acompanharem essa jornada de perto e sempre dispostos a me ouvir e aconselhar. Vocês são partes importantes dessa conquista.

Ao amigo Euler Rodrigues por mergulhar junto comigo nessa dissertação e ter toda disponibilidade, colaboração e por momentos únicos de compartilhamento de ideias. Obrigado por todo o incentivo.

Ao Prof. Dr. Silvânio Silvério Lopes da Costa por todos os ensinamentos, por compartilhar de experiências, momentos de descontração, coletas, trabalhos e aconselhamentos importantes nessa caminhada.

A Prof. Dra. Adnivia Santos Costa Monteiro pelos momentos de risadas, desabafos, acolhimento e aprendizado durante esse período que sempre serão lembrados.

A Filippe Sobral da Silva pela amizade, pelo apoio e as boas risadas ao longo desses últimos.

Aos amigos dos laboratórios LQA e LTMA por todo incentivo e auxílio durante essa jornada.

A FAPITEC pelo fomento e apoio a pesquisa desenvolvida no mestrado e a SEMARH/SE e CETESB pela disponibilidade dos dados.

Por fim, mas não menos importante, a toda minha família e amigos que mesmo distantes incentivaram e sonharam junto comigo no caminho até esse momento. Obrigado!

RESUMO GERAL

A qualidade da água é essencial para a preservação e continuidade dos ecossistemas e, é também, fundamental no desenvolvimento das atividades humanas. Sendo assim, identificar e entender os fenômenos que nesta ocorrem, principalmente, a degradação decorrente desse último fator, é crucial no estudo dos recursos hídricos e exige que se análise os parâmetros físicos, químicos e biológicos, bem como suas alterações provocadas pelas atividades antrópicas. No entanto, esse estudo não é tão fácil devido a um monitoramento inadequado dos corpos hídricos e a escassez de dados que é característico dos programas de monitoramento. No sentido de entender os impactos das atividades antrópicas na qualidade da água, o uso da estatística e, também, das técnicas de aprendizado de máquinas são essenciais e precisam ser mais disseminadas e incentivadas como ferramentas que auxiliam na tomada de decisão na gestão dos recursos hídricos. Dentre as técnicas de aprendizado de máquinas, neste trabalho foram utilizadas as Redes Neurais Artificiais, com auxílio das séries temporais, e o *Random Forest* que permitiram a previsão e a predição da concentração de clorofila-a, um dos principais indicadores do processo de eutrofização. Para predição e previsão da clorofila-a foram usadas as bases de dados da Companhia Ambiental do Estado de São Paulo (CETESB) e da *United State Geological Service* (USGS). A Rede Neural artificial exógena foi utilizada para previsão da clorofila-a e obteve-se concordância entre os valores preditos e mensurados tanto para o conjunto de dados de teste e de treinamento, conforme análise de métricas utilizando Erro quadrado médio (MSE) e a Raiz do Erro Quadrado Médio (RMSE). O *Random Forest* foi utilizado para a predição de clorofila-a em reservatórios do estado de Sergipe, e mesmo com uma base de dados menor, o modelo obteve bons resultados que poderiam ser melhores se houvesse mais dados disponíveis. Sendo assim, esses algoritmos apresentaram acurácia em seus resultados que podem permitir reduzir custos de análises laboratoriais e de mão de obra.

Palavras-chave: Aprendizado de Máquinas; Clorofila-a; Qualidade da Água.

ABSTRACT

Water quality is fundamental to the preservation and continuity of ecosystems and is also fundamental in the development of human activities, therefore identifying and understanding the phenomena that occurs in it, as the degradation resulting from this last factor, is fundamental in the study of water resources. In order for this study to attend and serve the diverse stakeholders in the use of water, identifying the physical, chemical and biological parameters and their changes caused by anthropic activities is essential. Nevertheless, this study is not so easy due to the lack of adequate monitoring of water bodies and the scarcity of data that is characteristic of monitoring programs. In this sense, in this crucial quest to understand the impacts of anthropogenic activities on water quality, the use of statistics as well as machine learning techniques are essential and also needed to be further disseminated and encouraged as tools that help decision makers. In this paper, the Artificial Neural Networks, with the aid of the time series, and Random Forest were used, which allowed the forecasting and prediction of the chlorophyll-a concentration, which in high concentrations characterizes a water body as eutrophic due to the excess of nutrients present in it. In this manner, it was possible forecasting chlorophyll-a concentration using Companhia Ambiental do Estado de São Paulo (CETESB) and the United State Geological Service (USGS database, Artificial Neural Networks was used, and results were obtained close to the values of test and train sets, according to the metric analysis using Means Square Error (MSE) and the Root Mean Square Error(RMSE).The Random Forest algorithm was used to predict the chlorophyll-a concentration in reservoirs in the state of Sergipe, and even with a smaller database, the model obtained good results that could be improved if more data were available. Therefore, these algorithms presented accuracy in their results that can allow reduce costs of laboratorial and labor analyzes.

Keywords: Machine Learning; Chlorophyll-a; Water Quality.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO GERAL

A escassez hídrica ocorrida nas últimas décadas acende a luz da necessidade de implantação de ações que possam monitorar e evitar problemas maiores com a pouca água existente com qualidade para consumo humano, de forma atender a todos os atores e a suas necessidades. Nesse sentido, a avaliação da qualidade da água insere-se como peça chave para que essa quantidade de água não se deteriore ainda mais e que medidas sejam tomadas enquanto ainda é possível. As avaliações dos parâmetros físicos, químicos e biológicos da água são, então, fundamentais para o processo de entrega de uma água de boa qualidade.

Para tal avaliação, identificar as características do corpo hídrico auxilia no entendimento dos fenômenos que ali ocorrem ou poderão ocorrer devido ao tipo de atividade antrópica desenvolvida nas bacias hidrográficas que impactam diretamente sobre os ecossistemas e a saúde humana. Essa identificação permite evitar problemas ambientais como a deterioração parcial ou completa do corpo hídrico, economia no tratamento e na distribuição de água e gastos medicinais com possíveis doenças que possuem a água como vetor de transmissão.

Dentre os problemas resultantes de atividades antrópicas, a eutrofização constitui um dos problemas mais graves que limita o uso da água. Este fenômeno vem sendo avaliado através de análises físico-químicas e microbiológicas que exigem uma visão macroscópica do sistema hídrico como um todo. No entanto, essas análises necessitam de periodicidade e de uma definição robusta de quais parâmetros ambientais devem ser mensurados. Dentre estes a concentração de clorofila-a como indicador do fenômeno de eutrofização, sendo esta indicadora de aporte excessivo de nutrientes advindo, por exemplo, de despejos de esgoto doméstico ou de fertilizantes agrícolas.

Ademais, entender e predizer variações de variáveis ambientais usando técnicas estatísticas e de aprendizagem de máquinas têm contribuído como ferramenta importante de tomada de decisão para os gestores, gerando benefícios como tempo de resposta relativamente menor frente a dispendiosas análises laboratoriais e custo de logística de coleta (PALMER *et al*., 2015; LOU *et al.* 2016; KELLER *et al*., 2018). No entanto, parte deste trabalho é prejudicado pela necessidade de uma quantidade de dados significativa que muitas vezes não está disponível, o que traria maior precisão a tais métodos.

Vários são os trabalhos que mostram essa relação fenomenológica entre a qualidade da água e os limites de concentrações de clorofila-a e de outros parâmetros ambientais. Zhang *et al*. (2015), Hollister *et al*. (2016), Li *et al*. (2017), Kovalenko *et al*. (2018) e Shoda *et al*. (2019) são trabalhos que merecem destaque quanto à avaliação desse fenômeno e abordam o uso de técnicas de aprendizagem de máquinas.

Zhang *et al.* (2015) avaliaram a qualidade da água no reservatório Yuqiao, na China, utilizando as redes neurais para predição (*forecasting*) de variáveis relacionadas ao fenômeno de eutrofização: temperatura, fósforo total e clorofila-a. O modelo mostrou-se capaz de predizer os níveis de eutrofização em até duas semanas o que auxiliaria a tomada de decisão dos órgãos gestores da região e contribuiu no monitoramento da floração algal deste reservatório. Hollister *et al.* (2016), buscando avaliar também o fenômeno de eutrofização, realizaram a modelagem utilizando *Random Forest* para a predição da clorofila-a para 1143 lagos dos EUA. A base de dados contém diversas variáveis ambientais, sendo assim, um estudo para identificação das variáveis mais importantes e seu impacto na presença ou não no modelo foi analisado pelos autores. Li *et al.* (2017) avaliaram o estado trófico do lago Poyang, na China, utilizando 11 parâmetros de qualidade da água, em 13 pontos entre os anos de 2008 e 2014. Os autores utilizaram a Análise de Componentes Principais (PCA) e o *Random Forest* para a observar a relação dos parâmetros de qualidade da água com a clorofila-a e para a predição de clorofila-a no corpo hídrico, respectivamente.

Kovalenko *et al.* (2018), ainda, referem-se a mudanças significativas em concentrações de diversos nutrientes e parâmetro de qualidade da água. O aumento continuo das concentrações dos compostos nitrogenados ocorreu devido a deposição de nitrato e o elevado tempo de residência do corpo hídrico. Por fim, Shoda *et al.* (2019), Shoda *et al.* (2019) buscaram identificar através de sete parâmetros de qualidade da água em 762 pontos de monitoramento nos EUA, de 47 diferentes agências, durante os anos de 2002 e 2012, como a qualidade da água se altera ao longo do tempo e as suas implicações a saúde humana e a vida aquática.

Neste contexto, como forma de compreender ou avaliar o cenário hídrico atual, visando sanar dificuldades com a disponibilidade de dados, a presente dissertação propõe dois artigos. O

primeiro artigo deste trabalho utiliza as Redes Neurais Artificiais com séries temporais aplicados a corpos hídricos nos EUA e no estado de São Paulo, afim de atingir uma melhor performance do algoritmo bem como possibilitar uma predição de clorofila-a, variável alvo associada ao aparecimento de águas decorrente de excesso de nutrientes em água, mais próxima da realidade. No segundo artigo, para análise dos reservatórios no Estado de Sergipe utilizou-se o *Random Forest* como algoritmo capaz de melhor performance e adaptação a escassez de dados para tais corpos hídricos.

Esse trabalho possui, então, através desses dois artigos, o objetivo de possibilitar o uso de ferramentas computacionais para auxiliar a gestão de um bem fundamental para o ser humano que é a água, avaliando sua qualidade por meio de variáveis ambientais na predição e previsão da concentração de clorofila-a.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Importância da Qualidade da Água

A qualidade da água vem se deteriorando ao longo dos anos e medidas devem ser tomadas para que esse bem tão importante para a existência da vida da Terra, não se torne ainda mais escasso. A análise da deterioração da qualidade da água é bem complexa e não linear, exigindo diversos estudos e grupos multidisciplinares para ampliar o entendimento e verificar caminhos para diminuir a vulnerabilidade dos recursos hídricos, atendendo a demanda da sociedade em seus diversos setores. Para monitorar eventos que ocorrem em um corpo hídrico é necessário conhecimento das possíveis fontes de contaminação, através do reconhecimento das atividades antrópicas desenvolvidas na bacia hidrográfica, regime hídrico, geologia e clima da região.

Neste contexto, os modelos propostos para avaliar corpos hídricos devem levar em consideração esses fatores em seu desenvolvimento. No desenvolvimento dos modelos é importante ressaltar outros aspectos como disponibilidade de dados para a construção dos mesmos e estes devem facilitar a incorporação de novos dados que atualizem o mesmo buscando aumentar sua performance. A aplicabilidade desses modelos deve abranger aplicações em mais de um corpo hídrico, o tornando ainda mais aceito (LOUCKS *et al.*, 2005; LOUCKS e VAN BEEK, 2017). Os modelos de uma forma geral, muitas vezes, são fortemente influenciados pela complexidade das variáveis físico-químicas e as suas interações com o corpo hídrico, fazendo com que a calibração dos modelos fique mais complexa. Além disso, a escolha dos modelos passa pela disponibilidade de dados para que este tenha acurácia e reprodutibilidade.

Os modelos quando propostos não tomam decisões por si só, eles indicam possíveis caminhos a serem traçados para os responsáveis pelas tomadas de decisão. Eles são mais uma ferramenta capaz de retratar a realidade dos corpos hídricos e as consequências, muitas vezes, das ações antrópicas inconsequentes que acarretam diversos problemas como o da eutrofização. Outra razão para os modelos não tomarem decisão, é que existem diversos atores interessados no uso dos corpos hídricos o que dificulta, por vezes, o consenso e o uso do modelo da forma mais correta, cabendo aos gestores a missão de equilibrar os interesses.

Um relatório do Instituto Nicholas da Universidade de Duke, EUA (2015) em parceria com o Instituto Aspen, chama a atenção para a necessidade de maior organização das bases de dados dos institutos de pesquisa em qualidade de água, visando uma melhor intepretação da base de dados pelos analistas e, consequentemente, um melhor diagnóstico do corpo hídrico. Existem hoje diversas formas de análise de águas, não somente em laboratórios, mas através do sensoriamento remoto e do uso de drones, e que toda essa *big data*, necessita ser melhor organizada e planejada de forma a diminuir os custos efetivamente, ou seja, diminuir o tempo para filtrar esses dados, já que cada agência possui um perfil de coleta de registro dos dados.

Ainda sobre o relatório, referido anteriormente, este apresenta a necessidade do uso das diversas tecnologias existentes para a análise de corpos hídricos para o benefício da população, pois, muitas vezes os dados são registrados e não são divulgados para atender interesses específicos de grupos da sociedade. Com o avanço das tecnologias disponíveis para acompanhamento de corpos hídricos em tempo real facilita a tomada de decisão a predição de parâmetros de qualidade da água e antecipação das consequências de contaminação do corpo hídrico, mitigando possíveis catástrofes. Esses avanços, por vezes, esbarram em protocolos rígidos que os órgãos responsáveis possuem para divulgar os dados.

No relatório do ano de 2017, o Instituto Aspen afirma que os sistemas de registro de dados qualidade da água vêm apresentando melhorias no compartilhamento de dados, mas muitas agências ainda esbarram na falta de um financiamento efetivo o que acaba ocasionando dados faltantes em alguns parâmetros. As agências precisam perceber que os registros corretos das análises facilitam seu trabalho e diminuem os custos, já que eles teriam a possibilidade de predição, a depender do método, em tempo real do comportamento do corpo hídrico, e em relação, as estações de monitoramento poderiam ser identificadas as que são mais representativas e interessantes para o programa, e as menos representativas serem retiradas. Os dados de qualidade são tão importantes que a existência de dados confiáveis pode reduzir custos no tratamento de água e ajudar na inovação de sistemas de tratamento mais efetivos. Nesse sentido, os autores do relatório defendem que a publicidade dos dados deveria ser uma das prioridades das agências responsáveis pelos registros.

Gardner *et al.* (2017) relatam que mesmo as agências concordando que é importante o compartilhamento das informações de qualidade da água e uso de técnicas computacionais mais avançadas, os governos ainda possuem dúvidas sobre os verdadeiros impactos que essas tecnologias podem trazer, por isso, muitas vezes não realizam o investimento devido. Sendo assim, os autores buscaram apresentar os custos-benefícios para os diversos setores da sociedade quando há informações hidrológicas, químicas e biológicas a respeito da água disponíveis e compartilhadas. Além disso, Gardner *et al.* (2017) sugerem que mais estudos devem ser feitos para que os impactos em cada setor sejam melhores representados e menos especulativos, chamando assim, mais atenção dos governos para essa lacuna no que diz respeito ao compartilhamento de dados de água.

Ainda nesse contexto, Jeuland *et al.* (2018) defendem que há a necessidade de compartilhamento de dados de hidrologia e qualidade da água por parte das agências, mas para isso é necessário apresentar os estudos de viabilidade econômica o que tornaria as propostas mais aceitáveis para os governos. Além disso, os estudos devem buscar mostrar claramente o impacto para o consumidor final dos diferentes setores para que o apoio destes seja maior. Para isso, deve ser calculado os custos da capacitação dos agentes envolvidos desde a coleta e análise de dados, e assim, apresentar os benefícios a curto, médio e longo prazo as agências interessadas. Neste sentido ainda, é necessário informar aos órgãos a necessidade desses dados para que as predições sejam mais acuradas e estes tenham consciência real da importância do compartilhamento de dados.

Neste contexto, diversos trabalhos mostram essa necessidade de se conhecer os parâmetros da qualidade da água para que um banco de dados robusto possa ser compartilhado por todos os atores envolvidos na gestão dos recursos hídricos, merecendo destaque o trabalho publicado recentemente por Shoda *et al.* (2019).

Shoda *et al.* (2019) buscaram identificar através de sete parâmetros de qualidade da água (amônia, cloreto, nitrato, sulfato, sólidos totais dissolvidos, nitrogênio total e fósforo total) em 762 pontos de monitoramento nos EUA, de 47 diferentes agências, durante os anos de 2002 e 2012, como a qualidade da água se altera ao longo do tempo e as suas implicações a saúde humana e a vida aquática. Foram identificadas que 30% dos valores estavam acima do limite exigido, e que

apenas 6 locais estavam com valores acima ou muito abaixo desde 2002. Tais resultados indicam que não houve nem uma melhora significativa na qualidade, nem uma piora, em relações aos valores limites que seriam desejados para tais parâmetros.

Em relação aos parâmetros de qualidade da água, como amônia, nitrato, sulfato, cloreto e sólidos totais dissolvidos, Shoda *et al.* (2019) observaram ainda que há valores com tendência a estarem abaixo dos limites da *Environmental Protection Agency* (EPA), e os locais com valores acima, não estavam com um aumento considerável. No entanto, fósforo total e nitrogênio total apresentaram tendências a estarem acima do limite, identificando contaminações crescentes nas bacias hidrográficas de atividades relacionadas ao uso da terra, como a agricultura, o que pode levar a uma eutrofização. Os resultados apontam que se a qualidade da água for mantida nesses ambientes durante um longo período de tempo, os pontos que estão acima dos limites de concentração tendem a diminuir, antes dos locais em que a concentração estão abaixo, aumentarem as concentrações.

2.2 Clorofila-a

Dentre os parâmetros ambientais amplamente usados para definição ou caracterização da qualidade da água, merece destaque a clorofila-a. Este parâmetro quando em altas concentrações está relacionado, principalmente, ao fenômeno de eutrofização causado pelo aumento de nutrientes em água. Esse trabalho buscou verificar também como as alterações de concentração de clorofila-a interfere em outras variáveis de qualidade da água e buscar métodos mais céleres de obtenção dos valores de concentração da mesma.

Liping e Binghui (2013) trabalharam na predição da concentração de clorofila-a no rio Daning, na China, em um reservatório da hidrelétrica das Três Gargantas. Estes autores utilizaram a Análise de Componentes Principais (PCA) e uma Regressão Linear Múltipla (MLR). Eles avaliaram os parâmetros para a predição da concentração de clorofila-a em dois grupos. O grupo A com Temperatura, pH, Sólidos Dissolvidos, Sólidos Totais Dissolvidos, Sólidos Suspensos, Oxigênio Dissolvido, DQO, Nitrogênio Total, Nitrogênio Dissolvido Total, Fósforo Total e Fósforo Dissolvido Total. O grupo B contém o grupo A com mais dois parâmetros velocidade superficial da água (Vs) e o tempo de residência estimado (t).

Liping e Binghui (2013) ainda identificaram na regressão do grupo A, realizada com base nas componentes principais dessas variáveis, que a principal causa da floração de algas e consequente aumento das concentrações de oxigênio, estava relacionada aos níveis altos de nitrogênio e fósforo advindos do sedimento do corpo hídrico. No grupo B, as concentrações de fósforo total e dissolvido obtiveram maiores *scores* que as concentrações de nitrogênio total e dissolvido quando considerado os efeitos hidrodinâmicos (Vs e t). Por fim, ao validarem cada um dos modelos os autores observaram que a regressão do modelo B obteve resultados preditos com maior precisão, mostrando, assim, o impacto da hidrelétrica Três Gargantas na concentração de clorofila-a, já que a sua presença diminui a velocidade do rio, diminuindo a mistura vertical das camadas de água, aumentando o tempo de residência, consequentemente, as chances de floramento algal.

Em um trabalho pioneiro, Rajaee e Boroumand (2015) aplicaram Redes Neurais Artificiais (ANN), funções *wavelets*, Algoritmo Genético e Vetores de Suporte para Regressão (SVR), Regressão Linear Multivariada (MLR) para a predição de concentrações de clorofila-a na baía de San Francisco, EUA. Para comparar os resultados dos modelos, dos valores preditos e reais, foram utilizados o coeficiente de eficiência do modelo de Nash-Sutcliffe (E) e a Raiz do Erro Médio Quadrado (RMSE). Séries Temporais de Clorofila-a foram computadas e avaliadas com o coeficiente de determinação R^2 para verificar qual seria o melhor ajuste para os dados de entrada do modelo.

Rajaee e Boroumand (2015) identificaram que a melhor performance de predição estava nas Redes Neurais utilizando funções *wavelet*, pois, obtiveram o menor valore de RMSE 1,58, contra 5,53 da WMLR, 4,51 da ANN sem a função de transformação de *wavelet*, e 4,72 do Algoritmo Genético combinado a SVR. Esse valor é justificado pela capacidade mais acurada das redes neurais e seu potencial de considerar características físicas do ambiente, muitas vezes desconhecidas, que influenciem na concentração de clorofila-a. Além disso, os autores testaram dados com 240 meses de monitoramento e 60 meses de monitoramento, observando que a menor quantidade de dados faz com que a predição dos valores não seja adequada devido ao treinamento não ser, portanto, o ideal. Por fim, os autores sugerem um trabalho com series temporais utilizando

outras variáveis ambientais como fosfato e nitrato, para auxiliar na predição de clorofila-a nos meses seguintes a base de dados utilizadas.

Tizro *et al.* (2014) avaliaram parâmetros de qualidade da água (Sólidos Totais Dissolvidos, Condutividade Elétrica, HCO_3^-, SO_4^{2-}, Mg^{2+}, Ca^{2+}, Na^+,pH, Razão de Adsorção de Sódio (RAS)) buscando predizer valores futuros com base nas séries temporais na bacia do Rio Hor Rood, no Irã. Os autores observaram uma tendência de aumento das concentrações de todos os parâmetros avaliados menos de Na^+,pH, SAR. A tendência de aumento da Condutividade Elétrica, SO_4^{2-}, Ca^{2+} e HCO_3^- indicam sinais de deterioração da qualidade da água.

Além disso, Tizro *et al.* (2014) ainda observaram na avaliação dos erros entre os valores reais e preditos um RMSE muito baixo para todas as variáveis analisada, mostrando que o modelo de séries temporais adotado é de bastante precisão e se adequou bem aos dados avaliados do rio Hor Rood. Já o coeficiente de determinação (R^2) foi maior que 0,66 para todas as variáveis, menos para o SO_4^{2-}.

Verma e Singh (2013) buscaram predizer DBO e DQO em uma região de mina de carvão em Jharkland, na Índia, utilizando parâmetros como pH, temperatura, óleo e graxa produzidos na região. A ideia era predizer os valores dessas variáveis que muitas vezes não possuem valores precisos nas análises laboratoriais. Neste trabalho, os autores observaram que o RMSE foi muito próximo a 10%, tanto para a predição de DBO, quanto de DQO, mostrando que o modelo possui uma certa acurácia nas suas predições, o que pode trazer economia na quantidade de análises realizadas desses dois parâmetros. Os autores ainda ressaltam que a obtenção de mais dados para teste e treino poderá contribuir para o aperfeiçoamento deste modelo.

Stachelek *et al.* (2018) realizaram estudo sobre a qualidade da água em 775 lagos dos EUA, buscando avaliar os níveis de eutrofização dos mesmos ao longo e a qualidade dos dados fornecidos pela *National Eutrophication Survey* (NES), os quais foram digitalizados. Os autores chamam a atenção pela forma de como é feita as mensurações em cada região do país e que havia um único lago que teve medição de nitrogênio total, os demais realizaram amostras de nitrogênio em suas formas orgânicas e inorgânicas e que alcalinidade foi um dos parâmetros que mais obtiveram registros.

Ainda sobre o trabalho de Stachelek *et al.* (2018) é importante ressaltar que os autores publicaram os scripts dos seus programas para que futuros trabalhos façam uso dessa base de dados, e possam, dessa forma, auxiliar a predição de variáveis complexas como o tempo de residência nos reservatórios. Os autores alertam para a necessidade de padronização nos registros para que estudos importantes como esses possam ser realizados com mais frequência e as análises possam ser mais eficazes.

Rocha Junior *et al.* (2018) estudaram o impacto das secas na redução dos volumes de reservatórios devido as mudanças climáticas ocorridas nos últimos anos e como isso impacta diretamente no aumento dos nutrientes nesses corpos hídricos. Esse aumento acarreta na eutrofização o que é muito prejudicial a tão sofrida escassez no semiárido brasileiro. Os reservatórios observados apresentaram um significativo aumento das concentrações de nitrogênio, fosforo, clorofila-a e turbidez e redução da transparência nos períodos secos o que mostrou que o nível de eutrofização tende a aumentar com a diminuição dos volumes de água. A redução desse volume está associada ao aumento da evaporação nos momentos de seca e seca extrema, e a falta de precipitação.

Os autores, Rocha Junior *et al.* (2018) ainda ressaltam que os diferentes ecossistemas podem responder de forma diferente a escassez de água devido ao clima e a morfologia, como o caso de Medeiros *et al.* (2016) que observaram uma diminuição da floração algal devido a diminuição da penetração de luz no reservatório em função da suspenção dos sedimentos o que aumentou a turbidez do corpo hídrico.

Sinha *et al.* (2017) realizaram projeções de como as mudanças nos regimes de precipitação devido às mudanças climáticas podem influenciar no volume de nitrogênio que impacta os rios dos EUA. Eles observaram que com possíveis projeções de aumento do uso da terra por agricultores, por exemplo, e o aumento das chuvas, a quantidade de nitrogênio que escorre para os corpos hídricos tende a aumentar, o que ocasionaria uma elevação considerável da eutrofização. Para contrabalancear esse aumento das chuvas seria necessária, por exemplo, uma redução de 33% da carga de nitrogênio utilizada na agricultura para que os impactos sejam atenuados.

Ainda sobre o trabalho de Sinha *et al.* (2017) foi observado a importância dos órgãos de gerenciamento das bacias em avaliar os cenários observados, pois, as consequências a qualidade da água podem ser catastróficas para os próximos anos, conforme os modelos de predição para mudanças climáticas e pode ocorrer ainda o aumento da hipoxia desses corpos hídricos.

Costa *et al.* (2018) realizaram um estudo para avaliar as publicações na linha de pesquisa da eutrofização ao redor do mundo. Eles apresentaram que os EUA é um dos países que mais tem pesquisas na área de eutrofização, juntamente com a China, e que o Brasil possui um mínimo de 20 publicações por ano na área. Os autores ressaltam que os maiores avanços na área acontecem quando a colaboração internacional e os colaboradores são de diferentes áreas do conhecimento. Ainda de acordo com Costa *et al.* (2018) a análise da eutrofização por modelos com series temporais tem sido mais frequente do que com análises espaciais.

Budria (2017) apresenta as oportunidades que espécies transmissoras de doença como vírus, bactérias e fungos e beneficiam-se do aporte de nutriente nos corpos hídricos, da depleção de oxigênio e a redução da transparência para se desenvolverem em corpos eutrofizados. As mudanças ocorridas no corpo hídrico devido a eutrofização alteram a dinâmica dos ecossistemas oferecendo a oportunidade para esses vetores de doenças. A transmissão de doenças ainda pode ocorrer devido a muitos animais se alimentarem de organismos que tiveram contato com águas eutrofizadas, e acabam carregando consigo essa contaminação, e que se tiverem contato com o homem, transmitirão doenças.

Anagnostou *et al.* (2017) apresentam os 9 modelos mais usados em *machine learning* na avaliação de corpos eutrofizados trazendo vantagens e desvantagens de cada um. O autor ainda ressalta a importância da disponibilidade dos dados para a construção dos modelos e a aplicabilidade de cada um globalmente. Ainda neste trabalho, ressalta-se a necessidade de mais modelos para auxiliar os órgãos gestores a compreender e tomar decisões quanto aos corpos hídricos, buscando sua restauração e preservação.

Zhang *et al.* (2015) avaliaram a qualidade da água no reservatório Yuqiao, na China, utilizando as redes neurais para predição (*forecasting*) de variáveis relacionadas ao fenômeno de eutrofização: temperatura, fósforo total e clorofila-a. O modelo mostrou-se capaz de predizer os

níveis de eutrofização em até duas semanas o que auxiliaria a tomada de decisão dos órgãos gestores da região e contribuiu no monitoramento da floração algal deste reservatório. As séries temporais elaboradas com as diversas variáveis medidas foram baseadas no valor atual e em dois atrasos (*lag time*), t, t- 1 e t-2, para prever um valor futuro (t+1). A avaliação desses atrasos é consequência, também, da inconsistência e escassez de dados.

No trabalho de Zhang *et al.* (2015) a precisão do modelo foi verificada pelo cálculo do MSE e do R^2, para cada uma das três variáveis avaliadas como *target*. Os valores das redes neurais séries temporais e da predição mostraram-se bastante precisos quando tais métricas foram avaliadas.

Bohn *et al.* (2018) avaliaram imagens de satélite e valores mensurados *in situ* para realizar uma regressão para observar a relação da profundidade do disco de Secchi e a concentração de clorofila-a em um lago na Argentina. Para filtrar os dados obtidos via satélite e mensuração in situ, os dados mensurados correspondiam a mais de 10 anos de monitoramento, e as imagens de satélite, portanto deveriam acompanhar as datas de mensuração com um erro de no máximo 3 dias. Os modelos preditos pelas imagens de satélite ficaram bem próximos dos valores mensurados em campo, mostrando que o uso do sensoriamento remoto contribui para a ampliação da base de dados de monitoramento ambiental e identificação de fenômeno de eutrofização por meio do processamento dessas imagens.

Bohn *et al.* (2018) ressaltam ainda a importância econômica de uso do sensoriamento remoto no custo das coletas que serão realizados e na mitigação de provável eutrofização do corpo hídrico. Além disso, os mesmos instigam a avaliação da eutrofização com mais variáveis que as utilizadas no trabalho.

Em mais um estudo que contribuía para a tomada de decisão em reservatórios e regiões estuarinas, Park *et al.* (2018) utilizaram *Support Vector Machine* (SVM) e redes neurais artificiais para predição de clorofila-a em dois reservatórios da Coréia do Sul, JAR e YSR. Os dados para esta predição são de 7 anos de monitoramento desses corpos hídricos dos seguintes parâmetros: fosfato, clorofila-a, amônia, nitrato, temperatura da água, velocidade do vento e radiação solar. Estes parâmetros foram as entradas para os dois modelos de *machine learning* já citados.

Anagnostou *et al.* (2017) observaram que SVM obteve resultados um pouco melhores para esse estudo que as redes neurais. Eles mostraram que os erros foram menores devido a sua menor complexidade em relação as redes neurais para o conjunto de dados avaliado. Além disso, observaram, através de análises de sensibilidade, que a variável que mais influenciou a predição de clorofila-a no reservatório JAR para ambos os métodos foi o fosfato, e no reservatório YSR, amônia para SVM e radiação solar nas redes neurais.

2.3 Machine Learning

Técnicas de aprendizagem de máquinas (*Machine Learning*) vêm sendo amplamente utilizada para avaliação da qualidade da água com diferentes objetivos e com o senso comum de se ter cada vez mais informações ou dados sobre os corpos hídricos no mundo.

As técnicas de aprendizagem de máquinas que serão utilizadas neste trabalho são redes neurais e *random forest*.

2.3.1 Redes neurais

Uma Rede Neural Artificial (RNA) é um modelo preditivo motivado pela forma como o cérebro funciona, com uma coleção de neurônios conectados. Esse arranjo de redes conectadas por neurônio utiliza funções matemáticas não-lineares com a capacidade computacional de aquisição e armazenamento para posteriormente entregar o produto desejado pelo supervisionador da rede (SANTOS, 2014; GRUS, 2015).

A primeira rede neural foi criada em 1943 pelo neurofisiologista McCulloch e o matemático Walter Pits. No entanto, a tecnologia da época não permitiu os avanços que hoje são referenciados as Redes Neurais. Nas últimas décadas, com o avanço computacional, áreas como medicina, engenharia, economia e direito vêm aplicando as redes neurais como auxílio para a solução de problemas (SANTOS, 2014; CHANG *et al.*, 2015).

As RNA possuem basicamente duas fases de processamento, aprendizagem e aplicação da rede. Na fase de aprendizagem utiliza-se os hiperparâmetros (parâmetros de ajuste da rede) para

que seja estabelecida as características da rede como peso de cada conexão entre entradas, neurônios e as camadas. Sendo assim, para alcançar o objetivo da RNA a etapa de ajuste dos pesos é de grande importância para o aprendizado da mesma através das conexões (sinapses) estabelecidas (GARCIA, 2012).

O ajuste dos pesos é realizado através do treinamento da rede que pode ser supervisionado ou não supervisionado. No treinamento supervisionado a rede necessita de um especialista nos dados da rede para informar entradas e a saída esperada. A RNA, por sua vez, aprende e busca otimizar os valores de saída e ajustar os pesos que a compõe. Já no treinamento não supervisionado, o algoritmo da RNA busca identificar padrões e classes de um conjunto de dados, desenvolvendo a habilidade de formar classes, através das características dos grupos de entrada, sem a necessidade de um supervisionado (HAYKIN, 2001; SHKURIN, 2015).

2.3.1.1 Topologia da RNA

A topologia da RNA refere-se a quantidade de entradas da rede, ao número de camadas ocultas (*hidden layers*), o número de neurônios em cada camada, tipos de conexões entre os neurônios, taxa de aprendizagem, número de épocas, sendo este último, número de ciclos para convergência dos dados.

Dentre os tipos de redes neurais, as redes de perceptrons multicamada (MLP) são as mais utilizadas em diversas áreas. Estas são conhecidas pela capacidade de aproximação universal em que uma rede com uma única camada intermediária é suficiente para aproximação uniforme, sendo um conjunto de treinamento significativo para que qualquer função contínua seja representada (CYBENKO, 1989; ZHANG *et al.*, 2017). Essa rede de retropropagação (*backpropagation*) de erros é supervisionada. Uma MLP (Figura 1) consiste de uma única camada de entrada, uma oculta e uma de saída, totalmente conectadas, ou seja, ela consiste de no mínimo três camadas. A camada de entrada com a alimentação dos neurônios pelas variáveis de entrada, as camadas ocultas que são ajustadas pelas funções de ativação e seus pesos e a camada de saída. Sendo assim, por consequência, um neurônio em qualquer camada da rede está conectado a todos os neurônios da camada anterior. O ajuste dos erros entre o valor predito e o real é feito pelo ajuste dos pesos da camada (NIROOBAKHSH *et al.*., 2012; GARCIA, 2012; SHKURIN, 2015).

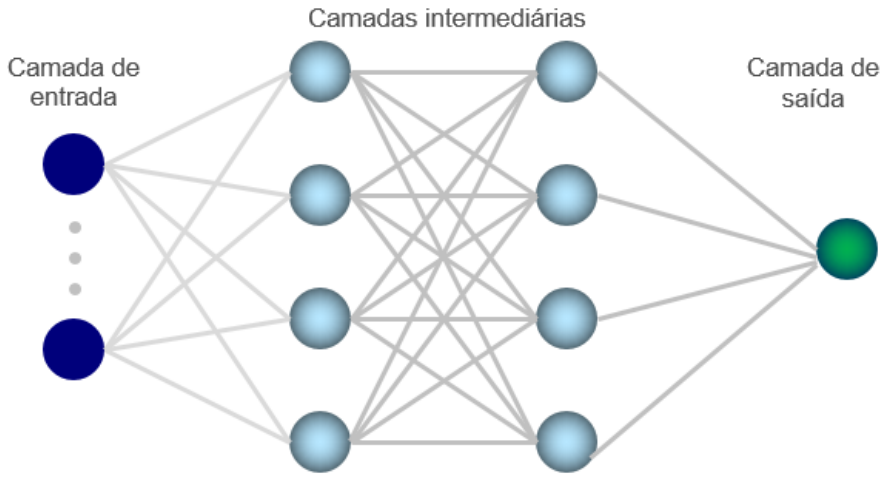


Figura 1 - Topologia da Rede Neural.
Fonte: Adaptada de Lacerda (2006)

Nas redes neurais, a técnica de parada antecipada (*Early Stopping*) é utilizada para que se evitar um *overfitting* do modelo desenvolvido. Em cada interação é verificada se há um melhoramento na rede, caso não seja identificado nenhuma melhora, o modelo é parado, evitando o *overfitting*. A quantidade de interações mínimas para que o modelo pare após a identificação é chamada de paciência (RASKUTTI *et al.*, 2014).

2.3.1.2 Modelo não linear auto regressivo com entrada exógena (RNAX)

Uma rede neural não linear RNAX é semelhante a MLP, entretanto, a rede é constantemente, retroalimentada pela saída com atrasos de tempo e uma entrada exógena com atrasos. Muitos trabalhos em engenharia sofrem com a falta de dados ou com a precisão dos mesmos, como os trabalhos que envolvem qualidade da água, dificultando o treinamento e a predição de uma variável. Para tal, a RNAX aliada às series temporais aparece com uma opção valiosa para sanar essas dificuldades (BOUSSAADA *et al.*, 2018; RUIZ *et al..*, 2016). A RNAX é eficiente para prever valores futuros de uma série temporal baseados nos valores passados dessas séries (variáveis de entrada) e os valores passados de outra série (variável de saída). (BOUSSAADA *et al.*, 2018; CHANG *et al.*, 2015).

Esse sistema não linear é representado matematicamente pela Equação 1.

$$z(t) = f[z(t-1),...z(t-dz);U(t)] \quad (1)$$

Em que U(t) e z(t) correspondem a entrada e a saída, respectivamente, no tempo. A função $f(.)$ é não linear e necessita ser aproximada para o aprendizado do algoritmo. A Figura 2 ilustra a rede RNAX, em que X(t) representa as variáveis na camada de entrada, Z(t) a realimentação com a série temporal da variável de saída.

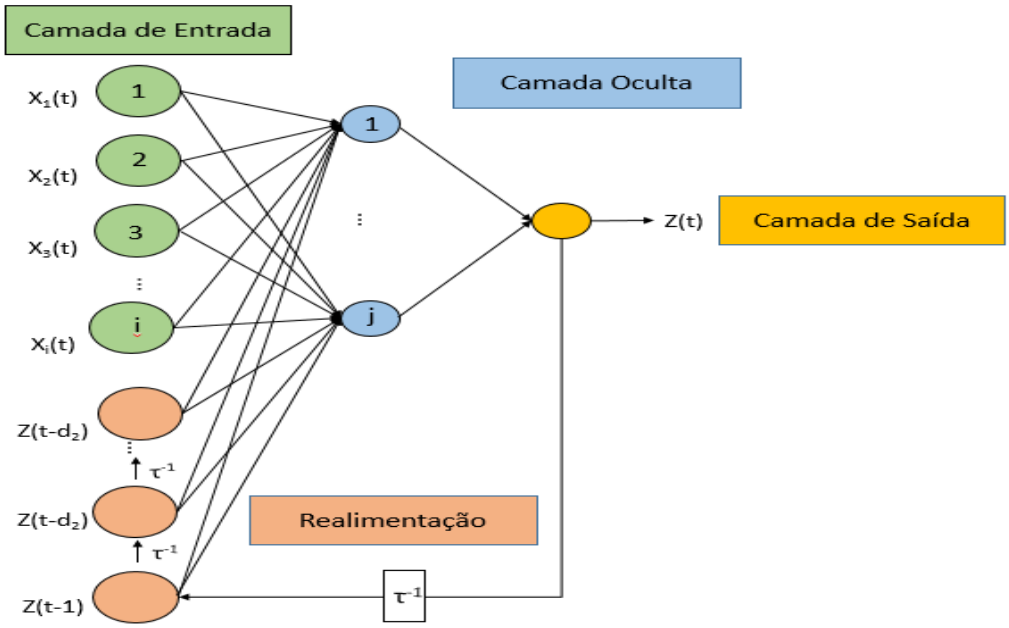


Figura 2- Topologia da rede RNAX
Fonte: Adaptada de Chang *et al.* (2015)

2.3.1.3 Funções de ativação (*Threshold function*)

As redes neurais são compostas de funções de ativação que possuem o papel de ativar ou desativar um neurônio, e para serem utilizadas os valores das variáveis devem ser normalizados de 0 a 1, ou -1 a 1 de acordo com a função definida. Alguns dos tipos de função de ativação são: linear, sigmoidal e tangencial (SANTOS, 2014; GRUS, 2015).

A função linear utilizada, geralmente nas camadas de saída da rede, está representada pela Equação 2.

$$f(x) = x \quad (2)$$

Já a função Sigmoidal (Equação 3) é representada por saídas entre -1 e 1 ou 0 a 1, essa última utilizada neste trabalho. Quando intervalo escolhido for entre 0 a 1, a função de ativação, pode ser referenciada na literatura também como tangencial, conforme Equações 3 e 4.

$$f(x) = \frac{1}{1 + e^{-px}} \tag{3}$$

$$f(x) = \frac{e^{px} - e^{-px}}{e^{px} + e^{-px}} = \tanh(px) \tag{4}$$

Além disso, há a função de ativação denominada de Unidade Linear Retificada (ReLU). Essa função usualmente, é bastante aplicada em redes neurais convolucionais, de fácil conversão e otimização (LECUN *et al.* ,2015). Seus erros de treinamento são geralmente, menores que a sigmoidal e tangencial. Sua descrição é bem simples, conforme Equação 5

$$\begin{aligned} &\mathrm{Re}\,LU(x) = \max\{0, x\} \\ &\mathrm{Re}\,LU'(x) = \begin{cases} 1, & \text{se x } \geq 0 \\ 0, & \text{caso contrário} \end{cases} \end{aligned} \tag{5}$$

2.3.1.4 Dropout

O *dropout* é uma técnica utilizada para evitar que o modelo da RNA não fique sobreajustado (*overfitting*) durante o processo de treinamento. Os neurônios que serão eliminados durante cada camada em treinamento são escolhidos de forma randômica. A rede, geralmente, é configurada para eliminar (*dropout*) as ligações entre neurônios e camadas em uma probabilidade 0,5 considerado o valor ideal. A utilização de *dropout* aumenta consideravelmente a quantidade de interações necessárias para que a rede acabe convergindo, no entanto, isso é compensado pela redução de épocas necessárias para que que o objetivo seja alcançado, conforme apresentado na Figura 3 (SRIVASTAVA *et al.*, 2014).

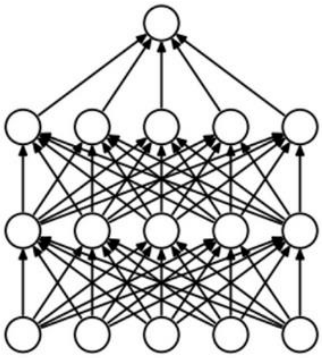
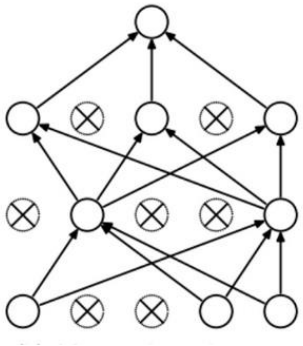
a)  b) 

Figura 3 - Topologia de Rede neural com Dropout
a) Rede Neural antes do dropout b) Rede Neural depois do dropout
Fonte: Srivastava *et al.* (2014)

2.3.1.5 Aplicações das Redes Neurais

Chou *et al.* (2018) utilizaram a base de dados de 20 reservatórios de Taiwan, no período de 1995-2016, utilizando diversas técnicas de *machine learning* para predição do Índice de Estado Trófico de Carlson (IET) com base em variáveis como temperatura, DBO, DQO, Sólidos Suspensos (SS), Amônia, que permaneceram como entrada, e mais 21 variáveis relacionadas a localização das amostras e as estações do ano. Os autores buscaram não utilizar nitrogênio, fósforo e clorofila-a para predição, pois, desejaram variáveis que fossem medidas com maior frequência no país, que não é o caso dessas. Assim, as variáveis utilizadas nos modelos híbridos e *ensembles* (utilizar modelos de *machine learning* compostos por mais de um método, que isoladamente são fracos) de acordo com a literatura citada pelos autores, se relacionam a contaminação das águas e especificamente, ao fenômeno de eutrofização.

Nesse sentido, dos modelos utilizados em diferentes softwares com a proposta dos autores, os modelos envolvendo as Redes Neurais, produziram os menores MAE e RMSE, apontando que estes poderiam ser adotados como alternativa mais econômica para avaliação da qualidade das águas, já que os parâmetros utilizados possuem uma maior facilidade de obtenção experimental. Chou *et al.* (2018) observaram que o software de melhor desempenho para os modelos utilizados foi o *IBM SPSS Modeler* com maior precisão dentre os demais.

Sprague *et al.* (2017) avaliaram a alimentação das bases de dados de qualidade de água nos Estados Unidos e ressaltaram alguns aspectos relevantes para que houvesse uma melhora na análise e distribuição dos dados. Os autores observaram que os registros dados não são uniformes nas agências o que dificulta a junção dos mesmos para usos secundários. Muitas vezes, obtém-se uma grande quantidade de informação dos dados, mas não há um padrão na forma de análise dos parâmetros nem nos seus registros como o uso das diversas formas químicas de fósforo e nitrogênio como um único registro para cada uma. Além disso, a quantidade de dados não alimentados pela falta de informação, por estar fora dos limites de detecção do equipamento, pela agência não querer informar os dados, ou ainda a incerteza de que se a mensuração aconteceu de fato, o que acaba tornando a análise de quem utiliza esses dados inconsistentes ou imprecisas.

Sprague *et al.* (2017) enfatizam, ainda, que as conclusões de um estudo que utiliza dados desse tipo podem relatar resultados diferentes quando analisados por vários grupos de estudo. Financeiramente, esse tipo de registro ainda é prejudicial pois muitas análises são realizadas repetidamente aumentando a ambiguidade dos dados. Sendo assim, estes autores sugerem uma harmonização nas formas de análises, nos registros dos dados e nas unidades utilizadas dos parâmetros facilitando o uso de usuários secundários que podem necessitar daqueles dados em seus trabalhos.

Rickwood e Carr (2009) utilizaram dados obtidos através do portal da Unesco GEMStsat para obter um índice de qualidade de água global baseado em equações e no índice já existente do *Canadian Council of Ministers of the Enfironment* (CCME). Os autores selecionaram os limites de concentração da Organização Mundial da Saúde (OMS) para definirem as faixas de qualidade do índice. A escolha dos parâmetros foi baseada nos que foram mais avaliados nos países, com um mínimo de quatro análises por ano, o que restringe o número de unidades de monitoramento inclusas, mas aumenta a consistência dos resultados.

Além disso, Rickwood e Carr (2009) sugerem que para trabalhos futuros a utilização de variáveis que não são incluídas pela OMS como comprometedoras à saúde humana se usadas sem nenhum tipo de tratamento avançado, mas se sabe que oxigênio dissolvido e DBO afetam o consumo e por isso, devem ser avaliadas para adição em índices futuros que trabalhem com essa

base de dados. Por fim, os mesmos relatam a validação do índice em uma bacia da Polônia e os resultados são compatíveis com os valores medidos das variáveis fornecidos pelo GEMStat

Peletz *et al.* (2018) avaliaram como é realizado o monitoramento em países da África subsaariana, buscando entender o que leva um programa de monitoramento ambiental obter sucesso ou não. Eles salientam que o poder aquisitivo de muitos países, e, consequentemente, o treinamento das equipes envolvidas, influencia na verificação precisa da qualidade da água. Muitos países nem mesmo sequer conseguem fazer o mínimo de amostras exigidos pela sua legislação. A dos dados utilizados no trabalho de Sprague *et al.* (2017) abrangeram 11 países, e tópicos com a capacidade das agências em divulgar os resultados para a população, capacidade de escolher locais de coleta e habilidade das equipes em interpretar e realizar os testes necessários para monitoramento necessitando mais fomento das autoridades.

Peletz *et al.* (2018) salientaram que agências de fornecimento de água, embora, tenham uma estrutura de funcionários mais restritas, estes são mais especializados que agências de vigilância e fiscalização, já que estes últimos não possuem, muitas vezes, o treinamento adequado, acesso a softwares e tecnologias mais avançadas, e o treinamento é breve ou inexistente.

Peletz *et al.* (2018) relacionaram a motivação dos funcionários ao tipo de liderança destes e que esta relação também impacta no resultado e análise das amostras de quantidade de água e que por isso algumas agências afirmam oferecer incentivos para tal. Problemas como transporte dos funcionários, feedback dos Ministérios responsáveis, falta de equipamento e burocracia para aquisição, também são mencionados. Os autores ainda alertam para a necessidade de investimento em tecnologia para análise de resultados, não apenas possuir dados, mas saber utilizá-los e interpreta-los, e monitoramento, e na especialização dos grupos responsáveis, bem como o desafio para os líderes de manter a motivação dos seus colaboradores mesmo quando há problemas estruturais.

Srebotnjak *et al.* (2012) buscaram construir um Índice de Qualidade de Água (IQA) único para o mundo. Os autores encontraram dificuldades na busca por dados, já que dados de qualidade de água com precisão e qualidade não são tão fáceis de se obter, muitas vezes devido ao sistema de monitoramento inadequado. Sendo assim, utilizaram a base de dados GEMStat da UNESCO

com mais de 100 países monitorados para propor o índice, que mesmo sendo uma base dados robusta, ainda assim, possui problemas como dados faltantes e não abranger todos os países do mundo, e contou ainda com contribuições de agências nacionais como as de Israel e Níger.

Baseada na opinião de especialistas, Srebotnjak *et al.* (2012) selecionaram cinco parâmetros (Oxigênio Dissolvido, Condutividade Elétrica, pH, Nitrogênio Total e Fósforo Total) que são indicadores ecológicos de depleção de oxigênio, eutrofização, acidificação e salinização. Além disso, tais parâmetros foram selecionados por estarem em uma quantidade de dados considerável na base utilizada, entretanto, métodos de imputação de valores foram utilizadas visando uma maior consistência do índice. Os valores limites das variáveis foram determinados com base nas diferentes legislações existentes ao redor do mundo. Srebotnjak *et al.* (2012) concluíram que a imputação de dados utilizada contribuiu para a precisão dos valores do índice proposto e que há a necessidade de um maior monitoramento da qualidade das águas visando melhorar tanto o entendimento dos fenômenos ambientais como também reduzir custo de tratamento de água e saneamento.

Rickwood e Carr (2009) utilizaram a base de dados da GEMStat, da UNESCO, buscando a validação de um índice de qualidade de águas. Os autores afirmam que existiu grande dificuldade para combinar os dados, pois, as diversas agências no mundo não medem as mesmas variáveis sempre. Para isso, um filtro foi aplicado buscando identificar as variáveis mais medidas nas agências que submeteram os dados, e que essa medição tenha ocorrido pelo menos quatro vezes em um ano. Neste sentido, os autores propuseram dois índices e afirmam que muitos países tiveram poucos dados registrados o que dificulta até mesmo a escolha de quais variáveis devem ser incluídas nos índices. Um índice foi baseado no guia de qualidade de água aceitabilidade e o outro nos limites definidos como saudável a saúde humana definida pela OMS.

Embora houvesse uma grande quantidade de variáveis com limites definidos pela OMS, há variáveis como poluentes orgânicos, associadas a água destinada para a agricultura que não estão com limites definidos no guia da OMS. Dessa forma, Rickwood e Carr (2009) sugerem que novos índices devem incluir esses parâmetros e que as medidas desses sejam mais frequentes pelos órgãos gestores, o que facilitaria a modelagem de novos índices.

Antonopoulos *et al.* (2001) avaliaram o rio Strymon localizado entre a Macedônia, Grécia e Bulgária com base nas séries temporais de nove parâmetros orgânicos, físicos, químicos e das descargas advindas de efluentes domésticos, industriais e da agricultura. A base de dado utilizada abrange um período de 28 anos. Os autores concluíram que as regressões utilizando valores de descargas diárias para parâmetros como nitrato, fósforo total, sulfato, cálcio e oxigênio dissolvido, tiveram melhor correlação com a vazão do rio, e que parâmetros como fósforo total e nitrato não possuem muita relação com sazonalidade, o que indica múltiplas fontes de descarga desses nutrientes.

Antonopoulos *et al.* (2001) concluíram que existe uma correlação inversa entre íons potássio e sódio, e com a condutividade elétrica, já que quanto menor a vazão menor será a diluição no corpo hídrico, consequentemente, aumentando a concentração dos primeiros ocasionando um aumento na condutividade elétrica, já que mais sais estarão presentes. O estudo de séries temporais desses parâmetros de qualidade de água deste trabalho ainda concluiu que as mudanças políticas ocorridas na Macedônia e na Bulgária na década de 80 impulsionaram a economia, aumentando a poluição do rio Strymon, enquanto, na década seguinte esse crescimento estagnou diminuindo a descarga de efluentes e aumentando as concentrações de oxigênio dissolvido, sódio e potássio presentes naturalmente no rio.

2.3.2 Séries Temporais

As séries temporais são um conjunto de dados de uma base, baseados em intervalos de tempo que pode ser definido em dias, meses ou anos. As séries temporais vêm sendo aplicadas em problemas de análises ambientais. Neste contexto, o objetivo da série temporal pode ser apresentado em duas frentes. A primeira buscando entender os mecanismos de uma dada variável em um período de tempo, e a segunda na predição de valores futuros de uma variável.

Neste sentido, uma série temporal é um conjunto de observações ordenadas no tempo (não necessariamente igualmente espaçadas), e que apresentam dependência serial (isto é, dependência entre instantes de tempo). Denota-se, então, uma série temporal Z_1, $.Z_2$,Z_3,...,Z_T, indicando por T o tamanho da série. O instante T indica o último instante da série analisada (WOOLDRIDGE, 2015; USP, 2018).

As séries temporais visam a análise e a modelagem de dados buscando verificar as características mais relevantes e suas possíveis relações com outras séries. Além disso, busca predizer uma série, a partir de valores passados da mesma ou até mesmo de outras variáveis, visando obter acurácia na predição de valores futuros. Dessa forma, sendo *n* um instante de previsão, ou horizonte de previsão, teremos que a previsão da série no instante T + n é denotada, por exemplo, da seguinte forma: $\hat{Z}_{T+n}$. (ABUDU *et al.*, 2010; WOOLDRIDGE, 2015; USP, 2018).

A base de dados deste trabalho consiste de séries temporais em tempos contínuos, no mesmo intervalo de tempo. Neste sentido, como as redes neurais desse estudo são usadas para interpretar relações não-lineares nos dados utilizados, o modelo de séries temporais utilizado foi o Modelo Autor regressivo Não-Linear (NAR). Esse modelo não-linear é utilizado para definir a entrada e saída em função do tempo, que é estimado de forma regressiva. Neste sentido, a Rede Neural Não-Linear Auto Regressiva (NAR-NN) é beneficiada tanto pelo NAR quanto pela ANN. Dessa forma, a equação geral da NAR-NN é apresentada na Equação 6.

$$y_t(t) = \sum_{j=1}^{m} w_{tj} y_{tj}(t-1) \qquad (6)$$

Em que $y_t(t)$ é o valor de saída da série $y_{tj}(t-1)$, é a entra da série, w_{tj} é o peso transferido do j-ésimo termo de entrada para o nodo da rede e m a quantidade de parâmetros de entrada (KHAN e SEE, 2016)

2.3.3 Random Forest

Random forest (RF) é um ensemble, uma combinação, de árvores de decisão e regressão que trabalham em paralelo, consideradas fracas isoladamente, e que juntas conseguem otimizar o modelo diminuindo desvios e variância (BREIMAN, 2001; AHMAD *et al.* 2018). Esse *ensemble*, combinação de modelos fracos que juntos conseguem otimizar o modelo, neste caso, árvores de decisão, pode ser utilizado tanto para regressões quanto para classificações de um conjunto de variáveis. Uma de suas vantagens é a não necessidade de normalização das variáveis consegue obter bons resultados com muitas variáveis de entrada e dificilmente tem resultados sobreajustados

(overfitting). O princípio do *Random Forest* é o sistema de voto (*voting*) em que cada árvore aponta quais as variáveis que mais ou menos influenciam na predição ou na regressão, tornando o modelo com menor probabilidade de erros. Segundo a definição de Breiman (2001) cada árvore seleciona de forma aleatória uma quantidade de variáveis de entrada para separação em nodo, para o cálculo da melhor forma de treino possível com tais variáveis. A árvore cresce usando a metodologia de Classificação e Regressão de Árvores (CART) para maximizar seu tamanho sem retirar as partes que não colaboram para a decisão. Matematicamente, esta técnica é descrita como um classificador formado por um conjunto de árvores de decisão $\{h(\mathbf{X}, Vk), k = 1,...\}$, em que *Vk* são vetores aleatórios independentes identicamente distribuídos e cada árvore possui voto para classe mais comum entre os vetores de entrada X (BREIMAN, 2001; RODRIGUEZ-GALIANO, 2015).

Dados de entrada não necessitam de uma normalização como em outras técnicas, já que os dados são selecionados de forma aleatória pela técnica *bootstrap*, estratégia de validação que a partir da formação de subconjuntos de teste e treinamento que permitem a inclusão de dados já sorteados em cada um dos subjconjuntos, ou seja, permite a reposição dos dados atribuindo a mesma probabilidade de sorteio a todos os dados do conjunto, buscando assim reduzir tendências e desvios associados ao modelo. Nessa técnica, parte dos dados, cerca de 1/3, não é utilizada para treinamento. Esse 1/3 servirá como comparação para os resultados da predição. A árvore é executada até atingir um critério de parada e, então, o seu valor é comparado aos dados que não fizeram parte dos dados de treinamento, pelo cálculo do erro denominado de *out-of-bag* (*oob score*). O modelo de *Random Forest* ainda associa para cada variável de entrada um valor de importância para que a redução do número de variáveis seja proposta visando otimizar o modelo (*feature importance*) (AHMAD,2018; DA SILVA *et al.*., 2017; IBAÑEZ, 2016; BREIMAN, 2001).

A Figura 4 ilustra o processo descrito sobre o *random forest*.

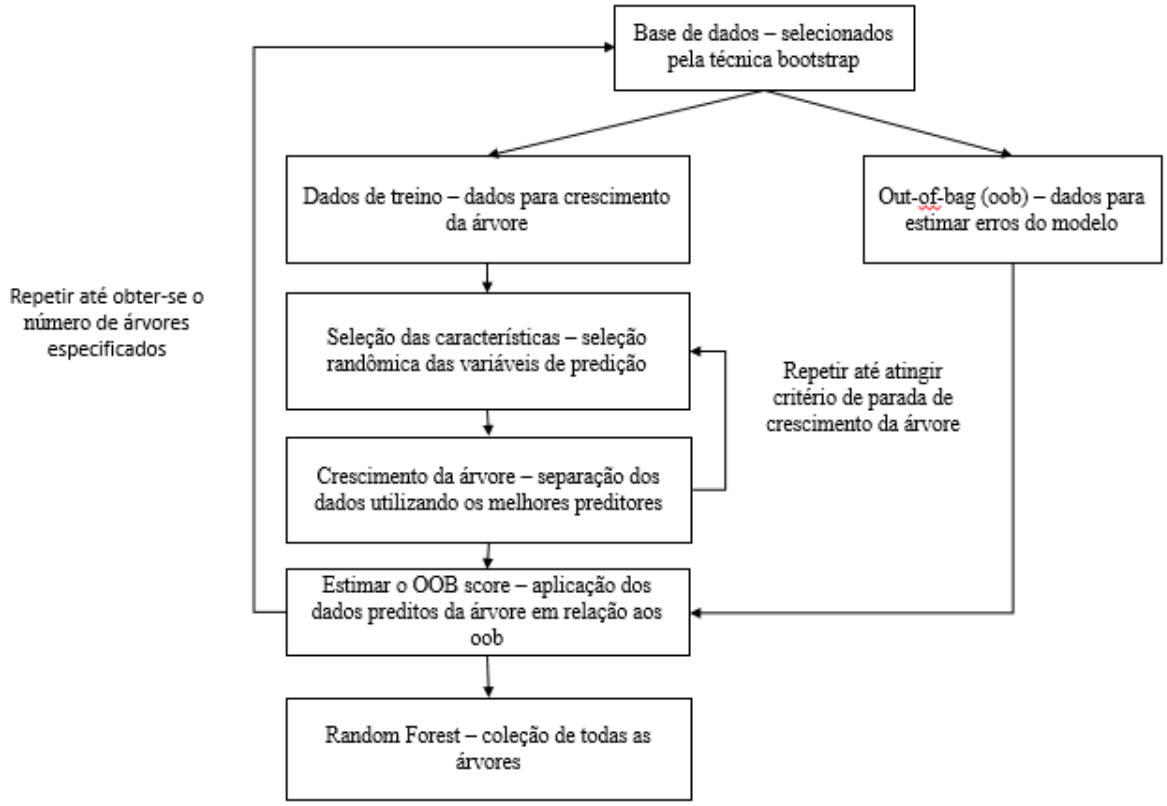


Figura 4 - Algoritmo Random Forest
Fonte: Adaptada de Barth (2013)

O cálculo do erro *out-of-bag* é apresentado na Equação 7:

$$Erro_{OOB} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}[y_i - x_i]^2 \qquad (7)$$

Em que yi é amostra presente no 2/3 dos dados de treinamento, xi são os valores presentes no 1/3 dos dados selecionados aleatoriamente como *out-of-bag*, n o número de árvores utilizadas para gerar o modelo.

2.3.3.1 Aplicações de Random Forest

Avila *et al.* (2018) utilizaram diferentes modelos estatísticos e de *machine learning* para predição de parâmetros da qualidade da água, especificamente de patógenos como E. *coli*. O corpo

de estudo foi o rio Oreti, na Nova Zelândia, e a base de dados utilizada abrangeu o período de verão entre os anos de 2006 a 2014. As outras variáveis analisadas foram a vazão do rio, os níveis de precipitação das últimas 24 e 48 horas, além do patógeno E. *coli.* O *Random Forest* e árvores de decisão foram utilizados, embora, não apresentaram os melhores resultados para o problema, o autor afirma que uma base de dados maior contribuiria para melhoria na performance deste modelo de regressão.

Hollister *et al.* (2016) realizaram a modelagem utilizando *Random Forest* para a predição da clorofila-a para 1143 lagos dos EUA. A base de dados contém diversas variáveis ambientais, sendo assim, um estudo para identificação das variáveis mais importantes e seu impacto na presença ou não no modelo foi analisado pelos autores. Os autores ressaltam a importância da análise do grau de eutrofização dos lagos, sendo este, um dos problemas que mais atinge esse tipo de corpo hídrico. O modelo utilizou 1000 árvores de regressão, e a avaliação do modelo foi feita pela raiz quadrada do erro médio (RMSE) e pelo valor do R ajustado.

Nesse sentido, Hollister *et al.* (2016) avaliaram variáveis medidas “in situ” e estimadas pelo Sistema da Informação Geográfica (SIG) e observaram que os modelos com todas as variáveis mensuradas em campo juntamente com as do GIS produziram um RMSE bem baixo, 0,08, o que mostra precisão do algoritmo e um R ajustado de 0,8. Na avaliação da importância de variáveis verificou-se uma melhoria do score quando apenas 6 das 20 variáveis foram utilizadas para predição: turbidez, fósforo total, nitrogênio total, elevação do lago, carbono orgânico total e relação N:P. Já a predição utilizando apenas as variáveis do GIS obteve um RMSE de 0,22 e um R ajustado de 0,48. Das 15 variáveis, seis foram consideradas mais importantes para a regressão: profundidade média do lago, ecorregião, percentual de área plantada, elevação, latitude, porcentagem de área verde. Os resultados mostraram que os modelos são capazes de contribuir na classificação dos corpos hídricos quanto a sua eutrofização devido a predição da clorofila-a.

Classificando lagos e reservatórios quanto ao seu nível trófico, Yuan e Pollard (2014), utilizando uma base de dados similar à de Hollister *et al.* (2016) buscaram classificar os lagos dos EUA utilizando *Random Forest* e observando as relações de fósforo total e clorofila-a. Modelos de predição de regressão com quantidades diferentes de variáveis foram avaliados, e o menor

RMSE apresentado foi o do modelo utilizando *Random Forest* que conseguiu reduzir o número de variáveis para predição frente aos outros modelos, principalmente, árvores de decisão. Sendo *Random Forest* uma combinação de árvores de decisão, os resultados obtidos estão de acordo ao adequados.

Utilizando a técnica de *random forest*, Yajima e Derot (2018) para a previsão da concentração de clorofila-a em um lago e um reservatório japonês com mais de 10 anos de registros com mais de 30 parâmetros de qualidade da água. Para a previsão da clorofila-a, os autores adotaram três atrasos nas séries temporais. Os autores observaram que o RMSE e tornou-se estável quando foram simuladas predições com mais de 200 árvores em cada um dos corpos hídricos. Para análise de importância nas simulações, as variáveis de entrada que foram consideradas mais importantes para duas estações do reservatório Urayama são DBO, DQO, razão entre Fósforo total e nitrogênio total e pH. Já para o lago Shinji, DQO e pH.

As variáveis de entrada com maior grau de importância no trabalho de Yajima e Derot (2018) estão ligadas diretamente ao fenômeno de eutrofização. Variações na concentração de DQO podem ser relacionadas ao aumento de fósforo e nitrogênio que estão ligadas ao aparecimento de algas em um corpo hídrico. Além disso, a importância das variáveis está ligada ao tipo de floração que tornará o corpo eutrofizado. Os autores ainda calcularam a correlação das variáveis e observaram que a turbidez possuía menor correlação com a clorofila-a, e foi apresentada por um dos rankings de importância das variáveis do *Random Forest* como maior correlação. A utilização dessa técnica de *machine learning* é capaz de apresentar relações que muitas vezes técnicas estatísticas básicas não conseguem informar. O algoritmo foi utilizado para a previsão e não para predição, o que pode explicar essas diferenças.

Li *et al.* (2017) avaliaram o estado trófico do lago Poyang, na China, utilizando 11 parâmetros de qualidade da água, em 13 pontos entre os anos de 2008 e 2014. Os autores utilizaram a Análise de Componentes Principais (PCA) e o *Random Forest* para a observar a relação dos parâmetros de qualidade da água com a clorofila-a e para a predição de clorofila-a no corpo hídrico, respectivamente. Os dados avaliados ao total chegam a mais de 3500. Em relação ao PCA, tanto na região ao nordeste do lago quanto na central, as 4 primeiras componentes foram responsáveis

por mais de 65% da variância total. Para a região central do lago a clorofila-a obteve melhores correlações com as outras variáveis, apresentando maiores scores na segunda componente juntamente, com oxigênio dissolvido e temperatura.

Para o modelo de *random forest*, Li *et al.* (2017) utilizaram 11 variáveis com entrada do modelo e a clorofila-a como variável alvo (*target*), tanto para a região central quanto para a região mais ao nordeste do lago. O modelo produziu melhor relação para os dados da região nordeste (R^2=0,61) do que para a região ao norte (R^2=0,46), o que pode ser associado a maior quantidade de dados da região nordeste do lago. Os autores ainda compararam o modelo a um simulado com redes neurais, e observaram menores resultados de RMSE para a predição com o *random forest*. O modelo *Random Forest* na análise de importância das variáveis estabelece a temperatura e nível da água como as mais importantes para o modelo da região nordeste, e para região sul, além desses dois, a profundidade do disco secchi também obteve maior importância para o modelo.

Li *et al.* (2018) utilizaram uma comparação da seleção de variáveis importantes utilizando mapas autoorganizaveis (SOM), um tipo de rede neural (KOHONEN), além do modelo de mínima redudancia e máxima relevância (mRMR) afim de comparar os modelos de predição *Random Forest* (RF) e *suport vector machine* (SVR), para dados do lago Baiyangdian, na China. Foram utilizadas 12 variáveis ambientais nos modelos para a predição de clorofila-a. Uma análise prévia da relação de Spearman, mostrou que nenhuma das variáveis apresentou correlação significativa com a clorofila-a tornando ainda mais necessária a aplicação de tais métodos de machine learning.

Li *et al.* (2018) ainda observaram que o ranking das variáveis para o *Random Forest* e para o mRMR foram diferentes. No entanto, em ambas DBO ficou em primeiro lugar. Além da DBO, o disco secchi ficou entre as primeiras em ambos. Para a regressão do SVR e *Random Forest*, os autores verificaram que os menores valores de RMSE e MAE para o *Random Forest*, mostrando a melhor predição deste método, principalmente, com valores muito altos da variável a ser predita. Observou-se quanto maior o número de variáveis melhor o modelo, principalmente, no que diz respeito do SVR. Li *et al.* (2018), constaram, também, que utilizar as 7 variáveis mais importantes atribuídas pelo *Random Forest* os valores de MAE e RMSE, tanto para o SVR quanto para o RF,

já não tinha uma melhora significativa, e que para o mRMR, os modelos tiveram melhor desempenho com 10 variáveis.

Vincenzi *et al.* (2011) aplicaram a técnica de ensemble de árvores de decisão, *random forest*, para avaliar a relação de plantas aquáticas com variáveis de avaliação da qualidade da água. O objetivo era relacionar o crescimento da *R. philippinarum* uma espécie de molusco, com fatores da qualidade da água. O modelo foi aplicado nas lagoas de Veneza, na Itália. A primeira análise feita foi a correlação quadrada de Spearman que demonstrou forte correlação da clorofila-a com a turbidez. No modelo de predição realizado pelo *Random Forest* as variáveis de maior importância para o desenvolvimento da *R. philippinarum*, foram salinidade, sedimentos e profundidade do corpo hídrico. Já as variáveis de menor importância foram oxigênio dissolvido, turbidez, tempo de residência, clorofila-a, temperatura e velocidade do fluxo de água.

O resultado obtido pelo *Random Forest* para o aparecimento da *R. philippinarum* está em conformidade com outros trabalhos utilizado na literatura, mostrando a excelente capacidade preditiva dessa técnica de machine learning. Os autores ainda citam esse modelo como forma de gestão das autoridades para controlar o aparecimento de algas no lago que pode prejudicar a cultura desse molusco na região.

Uma boa base de dados não somente em quantidade, mas em qualidade, faz total diferença nos resultados entregues pelos modelos de *machine learning*. Neste contexto, Kovalenko *et al.* (2018) avaliaram a qualidade dos grandes lagos canadenses com dados de 20 anos de monitoramento para identificar mudanças nestes corpos hídricos. Na predição de biomassa algal, no lago Michigan, utilizando *Random Forest* nitrogênio e sílica no Lago Michigan obtiveram maior importância para o modelo, cloro no Lago Ontario e no Lago Huron o ano de amostragem obteve maior importância. A relação do ano com o crescimento bentônico pode estar relacionada a contribuição de muitos parâmetros de qualidade de água ou outros fatores, segundo os autores. Já a sílica está relacionada a dinâmica das diatomáceas e que essas mudanças levam consequências a cadeia alimentar.

Kovalenko *et al.* (2018), ainda, referem-se a mudanças significativas em concentrações de diversos nutrientes e parâmetro de qualidade da água. O aumento continuo das concentrações dos

compostos nitrogenados ocorreu devido a deposição de nitrato e o elevado tempo de residência do corpo hídrico. A diminuição do processo de desnitrificação foi evidenciada no Lago Superior e menos evidente, mas não menos significante, no Lago Michigan. As alterações de fósforo nos Grandes Lagos também foram notadas, embora este possui baixa importância para o modelo de predição de fitoplancton, o que não quer dizer que não possua importância para o crescimento destes, mas as informações desse nutriente em específico foram reduzidas. As elevadas concentrações de cloro são significativas nos Lagos Huron, Michigan e Ontario, aumentado a salinidade destes. Essa salinização vem aumentado nos últimos 150 anos, devido ao escoamento no descongelamento das geleiras, poluição industrial e esgoto que acabam afetando as comunidades de plânctons e zooplanctons.

Odor e sabor são problemas que fazem com que pessoas duvidem da qualidade dos serviços de distribuição de água e que ocasionam aumento nos gastos para o tratamento desta água. Neste contexto, Kehoe *et al.* (2015) utilizaram séries temporais para avaliação do reservatório canadense Saskatchewan, no Canadá, utilizando *Random Forest* com dados de 24 anos. As variáveis preditoras para o odor foram clorofila-a, turbidez, fósforo total, temperatura e algumas espécies de algas produtoras de água. O modelo foi elaborado para prever o odor nas próximas 26 semanas. Os atrasos utilizados foram de 2, 12 e 26 semanas, sendo o atraso de 2 semanas o melhor desempenho, embora, o de 12 semanas tenha tido um desempenho semelhante.

Kehoe *et al.* (2015) utilizaram uma regressão linear, além, do modelo *Random Forest* que acabou tendo um desempenho inferior com um R^2 igual 0,52, e o do *Random Forest* igual a 0,71. Os maiores valores dos métodos analíticos de odor foram nos meses de agosto e esse período coincide com escoamento superficial, o degelo e aumento do aparecimento de fitoplanctons. As algas utilizadas para a predição, nos modelos simulados, obtiveram no geral maior importância frente as demais variáveis. Esse trabalho pode facilitar o planejamento das estações de tratamento de água para evitar maiores gastos, além de possibilitar, avisos a população de forma antecipada.

No reservatório Três Gargantas na China, onde se localiza a maior hidrelétrica do mundo, Zhang *et al.* (2017) buscaram identificar suscetibilidade de deslizamento de terra utilizando sensoriamento remoto com o método de classificação *Random Forest* e árvores de decisão. Esses

deslizamentos comprometem as propriedades da área e, por consequência, a vida da população que vive nessa região. Trinta e quatro fatores foram levados em consideração para a elaboração do mapa de suscetibilidade caracterizados pela topografia, geologia, hidrogeologia, cobertura do solo e gatilhos ambientais como chuva e intensidade sísmica.

Zhang *et al.* (2017) identificaram que poucas foram as áreas classificadas como altamente perigosas para deslizamento. Eles também enfatizam a melhor capacidade de predição do modelo *Random Forest* que modelos tradicionais de classificação abordados na literatura como as tradicionais árvores de decisão. Devido à grande quantidade de variáveis utilizadas na abordagem do modelo, a técnica *Random Forest* mostrou sua habilidade de predição com essa quantidade significativa. O modelo com 12 variáveis mostrou ter o melhor desempenho que o modelo com todas, com cerca de 1% a mais, sendo a sugestão dos autores para trabalhos futuros identificar qual seria o mínimo de variáveis importantes para o modelo possuir um desempenho ainda melhor. Os autores também sugerem este modelo como forma de gestão para engenheiros e órgãos gestores para minimizar os possíveis danos que os deslizamentos podem ocasionar.

Outra variável predita pelo *Random Forest* foi a contaminação fecal em corpos hídricos que servem de abastecimentos doméstico, em Bergen, na Noruega no trabalho de Mohammed *et al.* (2017). Eles utilizaram três variáveis dependentes (E. coli, bactéria coliforme, coliformes intestinais) e quatro variáveis independentes (condutividade, pH, cor e turbidez), em quatro estações do ano. O modelo *Random Forest* identificou cor e a estação outono com as de maior importância na predição da E. coli e bactéria coliforme. Na predição dos coliformes intestinais mostrou-se, no entanto, ser menor. Essa relação pode estar associada a circulação da água nessa época do ano, causando a presença de microrganismos na profundidade em que a água é retirada para abastecimento. As estações do ano mostraram significante efeito na predição de todas as variáveis dependentes e a turbidez mostrou ter efeito importante na predição da bactéria coliforme e nos coliformes intestinais.

Mohammed *et al.* (2017) quando simulada todas as oito variáveis independentes para a predição de cada uma das dependentes foram observados MSE elevados mostrando a necessidade de uma análise de importância das variáveis para redução do grupo de predição. Os resultados

mostraram que para predizer bactéria coliforme e coliformes intestinais, a condutividade e a cor foram as variáveis de melhor desempenho, enquanto a estação do outono foi a que mais ajudou a reduzir o MSE para a E. coli. confirmando assim a importância da análise de importância das variáveis na aplicação do *Random Forest*.

Ainda no contexto de avaliação da contaminação fecal em corpos hídricos, Roguet *et al.* (2018) avaliaram as bactérias *Clostridiales* e *Bacteroidales* presentes nas fezes de animais como cachorros, vacas, alces, porcos, nos EUA em sua maioria, nos anos de 2008 a 2016. Os esgotos também foram analisados em 17 cidades dos EUA nos anos de 2012 e 2013, em Reus na Espanha e em Salvador, no Brasil. Mais de 25 amostras no Lago Michigan foram coletadas, o que contribuiu no entendimento da contaminação fecal em diferentes locais do mundo.

O modelo *Random Forest* mostrou-se preciso no trabalho de Rogue *et al.* (2018) e os autores sugerem utiliza-lo com mais variáveis, para que a identificação do tipo de contaminante fecal seja ainda melhor, como identificar outros animais urbanos que possam contribuir com a contaminação fecal. As fontes que o modelo encontrou maior dificuldade para predizer foram as de gato e cachorro. Os autores ainda relatam a habilidade de predição do modelo levando em conta a velocidade com que se obtém muitas informações em curto intervalo de tempo, o que contribuiria para a gestão de patógenos.

Em corpos hídricos subterrâneos, Tesoriero *et al.* (2017) avaliaram a contaminação das águas em um conjunto de dados robusto utilizando *Random Forest*, no estado de Wisconsin, Estados Unidos. Dados de 1977 a 2016 foram avaliados em poços espalhados do estado, com o objetivo de identificar as fontes de maior contaminação do nitrato, arsênio e ferro na água subterrânea. Os autores definiram limites de concentração para avaliar a predição de modelo. Para locais com concentrações de nitrato $\geq$ 5 mg/L utilizaram doze variáveis para predição, apresentando como mais importante foram as plantações, mostrando que essas seriam a principal fonte do nitrato na água. Para as concentrações de nitrato $\geq$ 1 mg/L, nove variáveis foram avaliadas e o nível da camada de água foi a mais importante refletindo que menores concentrações de nitrato são esperadas nas regiões mais profundas, já que parte será consumida pelas reações redox com o O_2.

Ainda sobre o trabalho de Tesoriero *et al.* (2017), para os corpos hídricos com concentração de ferro ≥ 0,1 mg/L, oito variáveis foram analisadas, sendo a idade dos depósitos rochosos, recarga e profundidade da camada de água as mais importantes. Para o arsênio com concentrações ≥ 5µg/L a recarga obteve maior importância. Essa relação está relacionada as reações dos óxidos de ferros e sua presença nos sedimentos. Os autores ainda utilizaram modelos de regressão logística, mas o melhor desempenho foi do *Random Forest* mostrando obter melhor resultado frente as abordagens estatísticas tradicionais.

Em um contexto geral, as técnicas apresentadas neste capítulo da dissertação, quais sejam as redes neurais e *Random Forest* em conjunto com a estatística convencional estão sendo aplicadas com bastante sucesso para caracterizar, avaliar e até mesmo propor alternativas de gestão para os recursos hídricos. Sendo assim, nos capítulos 3 e 4 da presente dissertação serão apresentados dois artigos nos quais estão implementas as técnicas citadas.

REFERÊNCIAS

ABUDU, S.; KING, J. P.; BAWAZIR, A. S. Forecasting monthly streamflow of spring-summer runoff season in Rio Grande headwaters basin using stochastic hybrid modeling approach. **Journal of Hydrologic Engineering**, v. 16, n. 4, p. 384-390, 2010.

AHMAD, M. W.; REYNOLDS, J.; REZGUI, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. **Journal of Cleaner Production**, v. 203, p. 810-821, 2018.

ANAGNOSTOU, E.; GIANNI, A.; ZACHARIAS, I.. Ecological modeling and eutrophication—A review. **Natural Resource Modeling**, v. 30, n. 3, p. e12130, 2017.

ANTONOPOULOS, V.Z.; PAPAMICHAIL, D.M.; MITSIOU, K.A. Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. **Hydrology and Earth System Sciences Discussions**, *5*(4), pp.679-692, 2001.

AVILA, R.; HORN, B.; MORIARTY, E.; HODSON, R.; MOLTCHANOVA, E. Evaluating statistical model performance in water quality prediction. **Journal of environmental management**, *206*, 910-919, 2018.

BARTH, F. J. *Identificação de span utilizando random forest*. Faculdade BandTec e VAGAS Tecnologia, 2013. 11 p. Disponível em:<http://fbarth.net.br/materiais/docs/aula08.pdf>. Acesso em: 22/10/2018.

BOHN, V. Y.; CARMONA, F.; RIVAS, R.; LAGOMARSINO; L., DIOVISALVI, N.; ZAGARESE, H. E. Development of an empirical model for chlorophyll-a and Secchi Disk Depth estimation for a Pampean shallow lake (Argentina). **The Egyptian Journal of Remote Sensing and Space Science**, *21*(2), 183-191, 2018.

BOUSSAADA, Z.; CUREA, O.; REMACI, A.; CAMBLONG, H.; MRABET BELLAAJ, N. A. Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation. **Energie**s, *11*(3), 620, 2018.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BUDRIA, A. Beyond troubled waters: the influence of eutrophication on host–parasite interactions. **Functional Ecology**, v. 31, n. 7, p. 1348-1358, 2017.

CHANG, F. J.; TSAI, Y. H.; CHEN, P. A.; COYNEL, A.; VACHAUD, G. Modeling water quality in an urban river using hydrological factors–Data driven approaches. **Journal of environmental management**, *151*, 87-96,2015

CHOU, Jui-Sheng; HO, Chia-Chun; HOANG, Ha-Son. Determining quality of water in reservoir using machine learning. **Ecological Informatics**, v. 44, p. 57-75, 2018.

COSTA, J. A. D.; SOUZA, J. P. D.; TEIXEIRA, A. P.; NABOUT, J. C.; CARNEIRO, F. M. Eutrophication in aquatic ecosystems: a scientometric study. **Acta Limnologica Brasiliensia**, *30*, 2018.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of control, signals and systems**, v. 2, n. 4, p. 303-314, 1989.

DA SILVA, L.A.; PERES, S.M.; BOSCARIOLI, C.. Introdução à mineração de dados: com aplicações em R. **Elsevier Brasil**, 2017.

GARCIA, H. L. Desenvolvimento de Estratégias Para Utilização de Sistemas Inteligentes No Monitoramento da Qualidade da Água, Tese de Doutorado em Engenharia Química, Universidade Federal de Pernambuco, 2012.

GARDNER, J.; DOYLE, M.; PATTERSON, L. "Estimating the Value of Public Water Data." NI WP 17-05. Durham, NC: Duke University. Disponível em: http://nicholasinstitute.duke.edu/publications. Acesso em : 05/10/2018, 2017.

GRUS, J. **Data science from scratch: first principles with python**. " O'Reilly Media, Inc.", 2015.

HAYKIN, S., Redes Neurais, Princípios e Práticas, 2. ed., **Bookman**, Porto Alegre, 2001.

IBAÑEZ, M.M. Uso de redes neurais nebulosas e florestas aleatórias na classificação de imagens em um projeto de ciência cidadã. (Dissertação - Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais – São José dos Campos : INPE, 2016.

JAMI, M. S.; HUSAIN, I. A.; KABASHI, N. A.; ABDULLAH, N. Multiple inputs artificial neural network model for the prediction of wastewater treatment plant performance. **Australian Journal of Basic and Applied Sciences**

JEULAND, M.; HANSEN, K.; DOHERTY, H., EASTMAN, L.B.;TCHAMKINA, M. The economic impacts of water information systems: A systematic review. **Water Resources and Economics**, 2018.

KEHOE, M. J.; CHUN, K. P.; BAULCH, H. M. Who smells? Forecasting taste and odor in a drinking water reservoir. **Environmental science & technology**, v. 49, n. 18, p. 10984-10992, 2015.

KELLER, S.; MAIER, P.M.; RIESE, F.M.; NORRA, S.; HOLBACH, A.; BORSIG, N.; WILHELMS, A.; MOLDAENKE, C.; ZAAKE, A.; HINZ, S. Hyperspectral Data and Machine Learning for Estimating CDOM, Chlorophyll a, Diatoms, Green Algae and Turbidity. **International journal of environmental research and public health**, v. 15, n. 9, p. 1881, 2018.

KHAN, Y; SEE, C. S. Predicting and analyzing water quality using Machine Learning: A comprehensive model. In: **Systems, Applications and Technology Conference (LISAT), 2016 IEEE Long Island**. IEEE, 2016. p. 1-6.

KOVALENKO, K. E.; REAVIE, E. D.; BARBIERO, R. P.; BURLAKOVA, L. E.; KARATAYEV, A. Y.; RUDSTAM, L. G.; WATKINS, J. M. Patterns of long-term dynamics of aquatic communities and water quality parameters in the Great Lakes: Are they synchronized? **Journal of Great Lakes Research**, *44*(4), 660-669, 2018.

LACERDA, W.S. Notas de aula – Redes Neurais Artificiais, Ciência da Computação. Universidade Federal de Lavras, 2006.

LECUN, Y; BENGIO, Y; HINTON, G. Deep learning. **nature**, v. 521, n. 7553, p. 436, 2015.

LI, B.; YANG, G.;WAN, R.; HÖRMANN, G.; HUANG, J.; FOHRER, N.; ZHANG, L. Combining multivariate statistical techniques and random forests model to assess and diagnose the trophic status of Poyang Lake in China. **Ecological Indicators**, *83*, 74-83, 2017.

LI, X.; SHA, J.; WANG, Z.L. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. **Environmental Science and Pollution Research**, p. 1-11, 2018.

LI, X.; SHA, J.; WANG, Z.L. Chlorophyll-a prediction of lakes with different water quality patterns in China based on hybrid neural networks. **Water**, v. 9, n. 7, p. 524, 2017. (b)

LIPING, W.; BINGHUI,Z.. Prediction of chlorophyll-a in the Daning River of Three Gorges Reservoir by principal component scores in multiple linear regression models. **Water Science and Technology**, v. 67, n. 5, p. 1150-1158, 2013.

LOU, I.; HAN, B.; ZHANG, W. (Ed.). Advances in Monitoring and Modelling Algal Blooms in: Freshwater Reservoirs. **Springer**, 2016.

LOUCKS, D. P.; VAN BEEK, E. . Water resources systems planning and management: an introduction to methods, models and applications. **Springer**, 2017

LOUCKS, D. P.; VAN BEEK, E.; STEDINGER, J. R.; DIJKMAN, J. P.; VILLARS, M. T. Water resources systems planning and management: an introduction to methods, models and applications. Paris: Unesco, 2005.

MEDEIROS, L.C.; MATTOS, A.; LÜRLING, M.; BECKER, V. Is the future blue-green or brown? The effects of extreme events on phytoplankton dynamics in a semi-arid man-made lake. **Aquatic Ecology**, *49*(3), 293-307, 2015.

MOHAMMED, H.; HAMEED, I. A.; SEIDU, R. *Random Forest* tree for predicting fecal indicator organisms in drinking water supply. In: **Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on**. IEEE, 2017. p. 1-6.

NICHOLAS INSTITUTE - DUKE UNIVERSITY. Data Intelligence for 21st Century Water Management. A report from the 2015 Aspen-Nicholas Water Forum, 2015.

NIROOBAKHSH, M.; MUSAVI-JAHROMI, S.H.; MANSHOURI, M.; SEDGHI, H. Prediction of water quality parameter in Jajrood River basin: application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). **African Journal of Agricultural Research**, v. 7, n. 29, p. 4131-4139, 2012.

PALMER, S.C.; KUTSER, T.; HUNTER, P.D. Remote sensing of inland waters: Challenges, progress and future directions. **Remote Sens. Environ.** 2015, 157, 1–8.

PANDEY, D. S.; DAS, S.; PAN, I.; LEAHY, J. J.; KWAPINSKI, W. Artificial neural network based modelling approach for municipal solid waste gasification in a fluidized bed reactor. **Waste management**, *58*, 202-213, 2016.

PARK, Y.; CHO, K. H.; PARK, J.; CHA, S. M.; KIM, J. H. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. **Science of the Total Environment**, *502*, 31-41, 2015.

PELETZ, R.; KISIANGANI, J.; BONHAM, M.; RONOH, P.; DELAIRE, C.; KUMPEL, E.; MARKS, S.; KHUSH, R. Why do water quality monitoring programs succeed or fail? A qualitative comparative analysis of regulated testing systems in sub-Saharan Africa**. International journal of hygiene and environmental health**, 2018

RAJAEE, T; BOROUMAND, A. Forecasting of chlorophyll-a concentrations in South San Francisco Bay using five different models. **Applied Ocean Research**, v. 53, p. 208-217, 2015.

RASKUTTI, G.; WAINWRIGHT, M. J.; YU, B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. **The Journal of Machine Learning Research**, v. 15, n. 1, p. 335-366, 2014.

RICKWOOD, C.J.; CARR, G.M. Development and sensitivity analysis of a global drinking water quality index**. Environmental monitoring and assessment**, *156*(1-4), p.73, 2009.

ROCHA JUNIOR, C. A. N. D.; COSTA, M. R. A. D.; MENEZES, R. F.; ATTAYDE, J. L.; BECKER, V. Water volume reduction increases eutrophication risk in tropical semi-arid reservoirs. **Acta Limnologica Brasiliensia**, *30*, 2018,

RODRIGUEZ-GALIANO, V.; SANCHEZ-CASTILLO, M.; CHICA-OLMO, M.; CHICA-RIVAS, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, *71*, 804-818, 2015.

ROGUET, A.; EREN, A. M.; NEWTON, R. J.; MCLELLAN, S. L. Fecal source identification using random forest. **Microbiome**, v. 6, n. 1, p. 185, 2018.

RUIZ, L. G. B.; CUÉLLAR, M. P.; CALVO-FLORES; M. D.; JIMÉNEZ, M. D. C. P. An application of non-linear autoregressive neural networks to predict energy consumption in public buildings. **Energie***s*, *9*(9), 684, 2016.

SANTOS, T. F. M. D. Aplicação de séries temporais e redes neurais em um ambiente de computação em nuvem. Dissertação de Mestrado em Engenharia de Produção, Universidade Federal de Santa Maria, 2014.

SHKURIN, A. **Water Quality Analysis using machine learning algorithms**. Bachelor´s Thesis in Environmental Engineering. Mamk University Applied Science. Finland. 2015

SHODA, M. E.; SPRAGUE, L. A.; MURPHY, J. C.; RISKIN, M. L.. Water-quality trends in US rivers, 2002 to 2012: Relations to levels of concern. **Science of the Total Environment**, v. 650, p. 2314-2324, 2019

SINHA, E.; MICHALAK, A. M.; BALAJI, V. Eutrophication will increase during the 21st century as a result of precipitation changes. **Science**, v. 357, n. 6349, p. 405-408, 2017.

SPRAGUE, Lori A.; OELSNER, Gretchen P.; ARGUE, Denise M. Challenges with secondary use of multi-source water-quality data in the United States. **Water research**, v. 110, p. 252-261, 2017.

SREBOTNJAK, T.; CARR, G.; DE SHERBININ, A.; RICKWOOD, C. A global Water Quality Index and hot-deck imputation of missing data. **Ecological Indicators**, *17*, pp.108-119, 2012.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. **The Journal of Machine Learning Research**, *15*(1),p. 1929-1958, 2014

STACHELEK, J.; FORD; C.; KINCAID, D.; KING, K.; MILLER ,H.; NAGELKIRK, R. The National Eutrophication Survey: lake characteristics and historical nutrient concentrations. **Earth System Science Data**, *10*(1), 81-86, 2018.

TESORIERO, A. J.; GRONBERG, J. A.; JUCKEM, P. F.; MILLER, M. P.; AUSTIN, B. P. Predicting redox-sensitive contaminant concentrations in groundwater using *Random Forest* classification. **Water Resources Research**, *53*(8), 7316-7331, 2017.

THE ASPEN INSTITUTE. Internet of water: sharing and integrating water data for sustainability. A report from the aspen institute dialogue series on water data, 2017.

TIZRO, A. T.; GHASHGHAIE, M.; GEORGIOU, P.;VOUDOURIS, K. Time series analysis of water quality parameters. **Journal of Applied Research in Water and Wastewater**, *1*(1), 40-50, 2014

USP, Introdução às Séries Temporais. Disponível em <https://edisciplinas.usp.br/pluginfile.php/176834/mod_resource/content/1/Resumo%20Wooldridge%20-%20Introdu%C3%A7%C3%A3o%20%C3%A0s%20S%C3%A9ries%20Temporais.pdf> Acesso em: 27/08/2018.

VERMA, A.K.; SINGH, T.N. Prediction of water quality from simple field parameters. **Environmental earth sciences**, *69*(3), pp.821-829, 2013.

VINCENZI, S.; ZUCCHETTA, M.; FRANZOI, P.; PELLIZZATO, M.; PRANOVI, F.; DE LEO, G. A.; TORRICELLI, P. Application of a *Random Forest* algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. **Ecological Modelling**, *222*(8), 1471-1478, 2011.

WALSH, K. Dog river water quality: before and after precipitation. Disponível em: http://www.usouthal.edu/geography/fearn/480page/2011/11Walsh.pdf. Acessado em : 26/09/18.

WOOLDRIDGE, J. M. **Introductory econometrics: A modern approach**. Nelson Education, 2015.

YAJIMA, H.; DEROT, J. Application of the *Random Forest* model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. **Journal of Hydroinformatics**, v. 20, n. 1, p. 206-220, 2018.

YUAN, L. L.; POLLARD, A. I. Classifying lakes to improve precision of nutrient–chlorophyll relationships. **Freshwater Science**, v. 33, n. 4, p. 1184-1194, 2014.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B.; VINYALS, O. Understanding deep learning requires rethinking generalization. **arXiv preprint arXiv***:1611.03530*, 2016.

ZHANG, K.; WU, X.; NIU, R.; YANG, K.; ZHAO, L. The assessment of landslide susceptibility mapping using *Random Forest* and decision tree methods in the Three Gorges Reservoir area, China. **Environmental Earth Sciences**, *76*(11), 405, 2017.

ZHANG, Y.; HUANG, J. J.; CHEN, L.; QI, L. Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China. **Journal of Hydroinformatics**, *17*(4), 679-695, 2015.

3 REDES NEURAIS EXÓGENAS PARA PREVISÃO DE INDICADOR DE QUALIDADE DE ÁGUA

RESUMO

A aplicação de técnicas estatísticas e de aprendizado de máquina tem contribuído para a redução de custos nas coletas e, também, auxiliando a tomada de decisão. No entanto, para a aplicação destas técnicas é necessária uma base de dados bem estruturada e com disponibilidade significativa de dados para que a performance do algoritmo seja a mais eficaz possível. Neste trabalho, foram utilizadas duas bases de dados bem organizadas e que dispõem de uma gama grande de dados, USGS e CETESB, para que a concentração de clorofila-a fosse prevista. No entanto, mesmo essas bases sendo organizadas, foi necessário um tratamento de dados para retirada de alguns que, por exemplo, apresentavam periodicidades diferentes de análise e, também, que não tinham uma quantidade de parâmetros concomitantemente a existência da análise de clorofila-a. Sendo assim, buscando minimizar tais perdas foram utilizadas séries temporais dos parâmetros avaliados neste trabalho, bem como da clorofila-a para a previsão da concentração da mesma utilizando uma Rede Neural Artificial Exógena (RNAX). As métricas de análise de performance da RNAX, MAE e RMSE, 0,3731 e 0,5118, respectivamente, mostraram acurácia dos resultados obtidos quanto à previsão da concentração de clorofila-a. A partir desse excelente desempenho, foi realizada um estudo de caso utilizando os dados ambientais da bacia do rio Paraíba do Sul, em São Paulo. Para tal estudo, os valores de MAE 0,3974 e RMSE 0,5313, obtidos ratificaram a boa performance do modelo, indicando que é necessário um conjunto de dados ambientais consistente e organizado para construção e validação de modelos de redes neurais que auxiliem os gestores e analistas ambientais na tomada de decisão.

Palavras-chave: Redes Neurais; Qualidade da água; Previsão.

ABSTRACT

The application of statistics techniques and machine learning have contributed to the reduction of cost in the field analysis, and also helping in the decision making. However, to the application of these techniques, a well-structured database with significant data availability is necessary in order of the performance of the algorithm to be as efficient as possible. In this paper, two well-organized database that also has a with a large data range, USGS and CETESB, were used to forecasting the chlorophyll-a concentration. Nevertheless, even these databases well-structured, a data filtering was necessary for the removal of some samples, for example, presented different periodicities of analysis and, also, that they did not have a quantity of parameters concomitantly the presence of the analysis of chlorophyll-a. In order to minimize such losses, we used time series of the parameters evaluated in this paper, as well as chlorophyll-a to the forecasting of its concentration using an Exogenous Artificial Neural Network (NARX). The performance analysis metrics of NARX, MAE and RMSE, 0,3731 and 0,5118, respectively, showed accuracy of the results obtained for the prediction of chlorophyll-a concentration. From this excellent performance, a case study was carried out using the environmental data from the Paraíba do Sul river basin, in São Paulo. For this study, the MAE values 0,3974 and RMSE 0,5313, obtained confirmed the good performance of the model, indicating that a consistent and organized environmental data set is necessary for the construction and validation of neural network models that help managers and environmental analysts in decision making .

Keywords: Neural Network; Water Quality; Forecasting.

3.1 INTRODUÇÃO

A qualidade da água vem se deteriorando ao longo de décadas devido à urbanização, ao uso inadequado do solo, a processos industriais diversos, desenvolvimento econômico-científico e, principalmente, à falta de consciência ambiental da população. Sendo esse último ponto relacionado, principalmente, ao nível de desenvolvimento educacional da região. Além das atividades antropológicas, vale ressaltar que as variações sazonais que ocorrem em uma região acabam impactando sobre os parâmetros ambientais afetando diretamente a qualidade da água do corpo hídrico. Além disso é importante comentar que as características geológicas e morfológicas da bacia hidrográfica, também, devem ser avaliadas, pois interferem diretamente sobre parâmetros ambientais imprescindíveis para avaliação de um corpo hídrico.

Dentre os problemas ambientais que ocorrem em um corpo hídrico, a eutrofização resultante do aporte excessivo de nutrientes na água, é um dos problemas mais críticos. A presença de nutrientes acaba ocasionando a floração algal e, consequentemente, uma diminuição da concentração de oxigênio na água, já que boa parte deste acaba sendo consumida em processos oxidativos decorrentes da presença de algas, causando morte da fauna e flora do ambiente aquático. Outro parâmetro que é influenciado pelo excesso de nutrientes é a temperatura da água, já que a camada de algas presente acaba interferindo na penetração dos raios de sol na água prejudicando os processos naturais no corpo hídrico.

Além disso, no contexto da eutrofização, ambientes eutrofizados tendem a ter concentrações de carbonatos elevadas devido ao aumento dos processos fotossintéticos consequentes da floração algal, principalmente, em épocas de insolação elevada. Como consequência desse aumento de concentração verifica-se um impacto nas medidas de pH, já que a alcalinidade, também, é alterada.

Neste sentido, identificar se um corpo hídrico está passando por processo de eutrofização antrópica é fundamental. Dentre os diversos parâmetros ambientais, uma forma de determinar níveis de eutrofização é através da medida de concentração de clorofila-a na água. A concentração deste pigmento é afetada por parâmetros como concentrações nitrogênio e fósforo, geralmente advindos da lixiviação em áreas agrícolas ou de contaminação pontual ou difusa de esgotos, que

servem de nutrientes para as algas, além de outros parâmetros como disponibilidade de luz, turbidez e temperatura

Sendo assim, identificar formas de avaliar a variação temporal de parâmetros ambientais de forma precisa e financeiramente benéfica para os órgãos gestores é uma forma de evitar ou mitigar problemas de eutrofização, algumas vezes, irreversíveis naturalmente. Nos últimos anos técnicas estatísticas e de aprendizado de máquinas, como as Redes Neurais Artificiais, vêm ganhando destaque e interesse de diversos atores ambientais quanto à previsão da concentração de clorofila-a (RAJAEE e BOROUMAND, 2015; KHAN e SEE, 2016; CHOU et al., 2018). Sendo assim, modelos robustos, que buscam identificar formas de predição de uma variável ambiental, com base em séries temporais de variáveis ambientais tornaram-se ferramentas de grande contribuição científica.

No entanto, para utilização da inteligência artificial, em geral, é necessário o entendimento de como esta funciona e dos detalhes que fazem de um modelo preciso e consistente. Nesse sentido, para uma rede neural, por exemplo, é necessária uma quantidade de dados significativa para que o modelo aprenda e consiga prever valores que estejam na menor faixa de erro possível (CHANG et al., 2015; ZHANG et al., 2015). Em avaliações ambientais, um dos grandes problemas é justamente esta disponibilidade de dados e de programas de monitoramento efetivos que possam embasar um estudo com a melhor relação custo-benefício para os atores envolvidos e que este consiga retratar a realidade ambiental. Dessa maneira, uma das alternativas propostas é buscar redes de monitoramento que já possuem uma base de dados significativa e organizada para que o melhor desempenho do modelo seja atingido.

Neste sentido, o presente trabalho objetivou realizar a previsão da concentração da mesma utilizando uma Rede Neural Artificial Exógena (RNAX), alimentada por séries temporais de variáveis ambientais, a partir de dados de redes de monitoramento da qualidade de água.

3.2 METODOLOGIA

Considerando que a análise fenomenológica dos parâmetros ambientais independe do local, ou seja, se determinado parâmetro sofre uma variação de aumento ou diminuição o resultado será o mesmo, pois as interações são determinísticas, correlacionadas em relação ao tipo de

parâmetro e não seu local de coleta. Ainda mais explicitamente, por exemplo, o aumento das concentrações de Nitrogênio e Fósforo, seja aqui no Brasil ou na África, a consequência ambiental será a mesma, uma indicação da ocorrência do fenômeno de eutrofização, mesmo que os níveis de eutrofização dos corpos hídricos não sejam dependentes somente destas concentrações (RICKWOOD e CARR, 2009; SREBOTNJAK *et al*., 2012; SINHA *et al*., 2017).

Neste contexto, o presente trabalho obedeceu a uma lógica matemática no sentido de que um resultado mais preciso fosse obtido a partir de um maior número de dados. Em teoria, quanto mais exemplos se tem de um evento, melhor a capacidade preditiva de um modelo matemático e das técnicas de *machine learning*. Sendo assim, o desenvolvimento do modelo de Rede Neural para predição de clorofila-a foi organizado de acordo com o fluxograma presente na Figura 5.

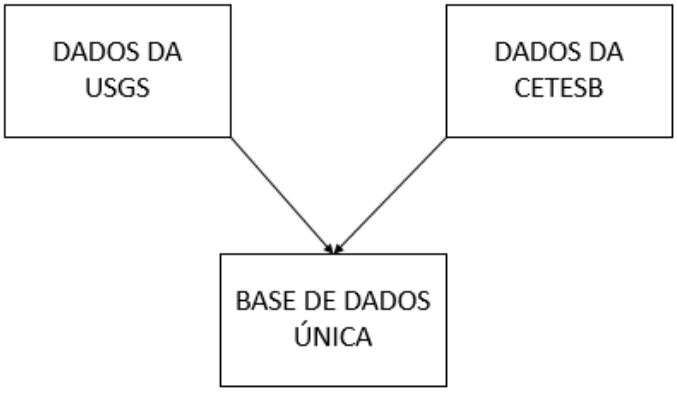


Figura 5 - Fluxograma das bases de dados

É importante informar que em uma primeira etapa foram solicitados dados de vários estados brasileiros aos órgãos responsáveis pela gestão dos recursos hídricos e a própria Agência Nacional de Águas (ANA). Desta solicitação, a melhor resposta foi advinda da Companhia de Ambiental do Estado de São Paulo (CETESB). Algumas outras agências que responderam a solicitação possuíam problemas quanto à uniformidade dos parâmetros mensurados e quantidade de dados disponíveis.

Sendo assim, foram solicitados dados dos EUA, África do Sul e Peru. No entanto, apenas as bases de dados advinda da *Environmental Protection Environment* (EPA) e da *United States*

Geological Survey (USGS) apresentavam uniformidade de dados e um programa de monitoramento das águas mais longevo.

Neste trabalho, utilizou-se de 36 parâmetros advindos das 9 variáveis selecionadas com suas respectivas séries temporais com três camadas ocultas com 300 neurônios utilizando-se de 1200 épocas para treinamento, o valor de paciência foi definido como 60 épocas, ou seja, épocas mínimas antes da rede finalizar seu treinamento. Além disso, o número de épocas necessárias para o fim do treinamento da rede foi igual a 275. A rede possui múltiplas entradas e uma única saída, esse arranjo é conhecido como MISO (*Multiple Input Single Output*) (JAMI *et al.*, 2012; PANDEY *et al.*, 2016).

A Figura 6 mostra a topologia simplificada da RNAX.

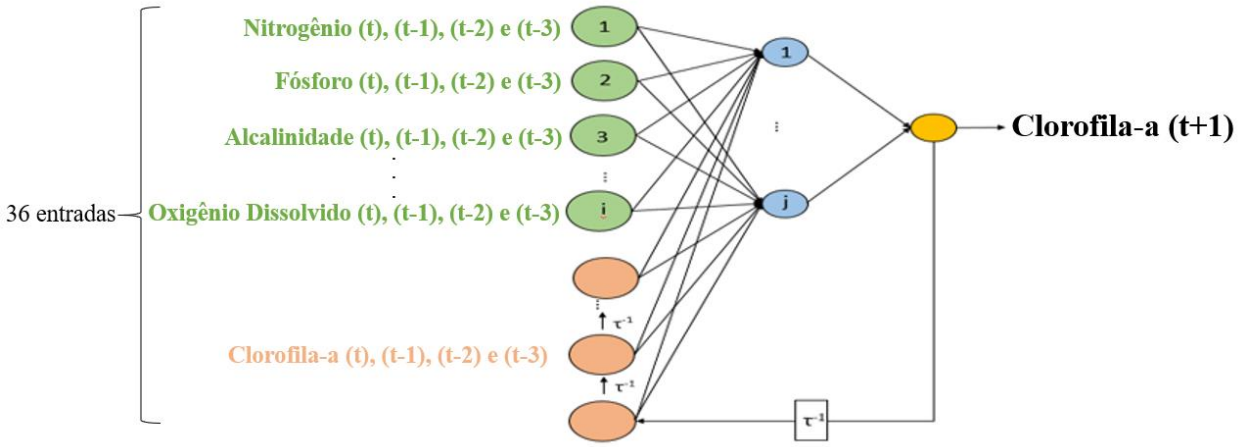


Figura 6 – Topologia da RNAX

3.2.1 Organização da base de dados

Para atender as necessidades da aplicação de métodos de aprendizagem de máquinas, buscou-se adquirir a maior série de dados possível e ao mesmo tempo condizente com o objetivo deste trabalho. Neste sentido, foram selecionados dados ambientais do Serviço Geológico dos Estados Unidos (USGS, 2018) de 9 estados americanos (Califórnia, Florida, Arizona, Missouri, Mississipi, Geórgia, Alabama, Texas, Louisiana). Esses estados foram escolhidos devido a sua posição geográfica mais ao sul dos EUA, ou seja, estados que apresentam condições climáticas

mais semelhantes e que as variações dos parâmetros não estivessem distantes das que acontecem aqui no Brasil. Neste sentido, os dados correspondem a um monitoramento de aproximadamente 25 anos, compreendendo 01/01/1993 a 01/07/2018. Além disso, para complementar a base de dados, adicionou-se dados nacionais da Companhia Ambiental do Estado de São Paulo (CETESB, 2018) e que foram obtidos dados de monitoramento hídrico ambiental dos anos de 2008 a 2017.

Após a seleção desses dois conjuntos de dados, em uma segunda etapa, utilizando a linguagem de programação Python no software Jupyter, identificou-se quais os parâmetros que mais foram analisados no monitoramento e foram retiradas as células vazias, o que reduziu a quantidade de dados. O passo seguinte, então, foi a identificação de um conjunto de parâmetros mais frequentes em todas as localidades afim de equalizar as características avaliadas para a predição de clorofila-a, assim, os parâmetros selecionados foram: Sólido Dissolvido Total (STD), Alcalinidade Total, Oxigênio Dissolvido (OD), Turbidez, Nitrogênio Kjedahl Total, Fósforo Total, Temperatura da Água, pH e Clorofila-a. A organização dos dados da CETESB assemelha-se aos dados da USGS, no que diz respeito aos parâmetros analisados e nos registros de campo, o que facilitou essa identificação, bem como todas as etapas seguintes deste trabalho. Após o filtro restaram dados da CETESB e de 4 estados dos EUA (Florida, Alabama, Arizona, Geórgia).

Uma estatística descritiva do conjunto de dados resultante foi realizada de forma a eliminar valores inconsistentes fenomenologicamente *(outliers*), como no caso da temperatura registrada de 1000 ºC, e pH de 20, que indicam erro na alimentação da base de dados. Além disso, células com valores nulos foram identificadas no *dataset* e então retiradas.

A padronização de sistemas de monitoramento no que diz respeito ao tipo de análise realizada, os registros de campo que são apresentados, facilita o trabalho das equipes responsáveis bem como a dos analistas ao longo dos anos. Desta forma, os resultados apresentados bem como o diagnóstico do que ocorre em um corpo hídrico fica muito mais pautado em dados concretos do que em suposições.

3.2.2 Séries temporais

Para a predição da variável clorofila-a nesse estudo utilizou- da seleção de variáveis que a influenciam (Sólido Dissolvido Total (STD), Alcalinidade Total (AT), Oxigênio Dissolvido (OD),

Turbidez(Tdz), Nitrogênio Kjedahl Total (NT), Fósforo Total (FT), Temperatura da Água (T), pH), baseados no *dataset* CETESB e USGS. Dessa forma, diferentes combinações entre essas variáveis foram avaliadas para predição do *target* (parâmetro alvo) clorofila-a no tempo t. Além disso, era necessário que cada ponto selecionado obtivesse no mínimo 3 amostras completas com tais variáveis para que houvesse a predição. Sendo assim era necessário x(t), x(t-1), x(t-2), para prever x(t+1).

Os *datasets* foram divididos em 80% dos dados para treinamento e 20% para teste em cada uma das localidades analisadas pelo modelo. O método *holdout* consiste nessa divisão dos dados nas categorias de teste e treinamento, e é aplicado amplamente nas diversas técnicas de *machine learning*, incluindo as RNA.

3.2.3 Análise de Performance de um Modelo

Com o objetivo de avaliar os modelos propostos em sua capacidade de predição foram calculados dois tipos de análise de performance: RMSE e MAE. O primeiro é o cálculo do valor quadrado médio (RMSE) entre o valor real e o predito pelo modelo, e o segundo é o valor médio do erro absoluto (MAE) (NODOUSHAN, 2018; MONTEIRO e COSTA, 2018). Essas métricas de performance são as mais utilizadas para avaliar a acurácia de um modelo, quanto mais próximos de zero os valores de RMSE e MAE maior a capacidade de predição do modelo. As Equações 8 e 9 apresentam o cálculo do MAE e do RMSE, respectivamente, sendo que n é o número de dados referente a variável de saída, o *target*.

$$\text{MAE}=\frac{\sum_{i=1}^{n}\left|y_{i(observado)}-y_{i(predito)}\right|}{n} \tag{8}$$

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_{i(observado)}-y_{i(predito)})^2}{n}} \tag{9}$$

Sendo: $y_{i(observado)}$=valor da variável mensurado, $y_{i(predito)}$=valor da variável predito pelo modelo, n=número de amostras.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Estatística Descritiva

Na análise da base de dados gerada pela combinação dos dados da USGS/EPA e CETESB, foi construída a rede com 5649 linhas. Os parâmetros analisados são apresentados nas Tabelas 1 e 2, juntamente com a estatística descritiva.

Observa-se que os valores de temperatura possuem uma elevada diferença com valores de 12,3 ^{o}C no rio Mobile, no Alabama, próxima captação de água da empresa responsável pelo tratamento MAWSS, no período do Inverno do hemisfério Norte, e no Rio Sorocamirim, monitorado pela CETESB, próximo ao ponto de captação na cidade de São Roque, no período de inverno do hemisfério Sul. Com o maior valor de temperatura 36,2 ^{o}C, destaca-se o rio Mogi-Guaçu, que segundo o relatório da CETESB (2015), possuía o segundo maior número de mortandade de peixes no estado de São Paulo, parte disso, devido a industrialização na bacia do rio Mogi-Guaçu.

Em relação aos sólidos totais dissolvidos (TDS), a Organização Mundial de Saúde (OMS) limita a concentração a um valor menor ou igual a 600 mg/L como aceitável para água utilizado para consumo humano. Na base de dados 50 a 75% das estações analisadas possuem valores de TDS entre 100 mg/L e 160 mg/L. No entanto, valores acima de 7000 mg/L foram encontrados em estações localizadas no estado do Alabama. As estações próximas à confluência Rio Dog e Halls Mill Creek apresentam um histórico de elevada concentração de sedimentos, conforme relatado por Cook e Moss (2012). O rio Dog recebe uma grande quantidade de sedimentos advindo de seus afluentes provocada pela erosão na região da cidade de Mobile, Alabama. A construção de canais para captar água demandada pela zona urbana, também, contribui para o aumento dos sedimentos, bem como o escoamento durante as chuvas que carreia partículas para o leito do rio conforme, também, comentam WALSH (2011). O desmatamento e o avanço da urbanização contribuem para o aumento dos sólidos nos afluentes do rio Dog, que acaba sofrendo com a acumulação em seu leito.

Na estação localizada na Florida, na região estuarina do rio Suwannee, obteve-se um dos maiores valores de TDS, 8310 mg/L, no ano de 2008, sendo a água imprópria para o consumo.

Conforme relatado por Balles *et al.* (2006) há muitos anos este rio que desagua no Golfo do México sofre com uma crescente salinização, apresentado, portanto, uma quantidade significativa de sólidos presentes na água, que são decorrentes de fenômenos como tempestades e furacões que acabam contribuindo na deposição de sedimentos, e da influência do próprio Golfo do México.

A estação da CETESB em Itanhaém (SP) apresentou os valores máximos da base de dados para TDS, no ano de 2012, no qual o relatório da CETESB (2012) o relaciona há um aumento de sedimentos devido as atividades antrópicas desenvolvidas na cidade. Esse mesmo ponto já registrou floração algal no ponto de análise, com um estado trófico classificado como Hipereutrófico no ano de 2009.

Tabela 1- Resumo estatístico dos parâmetros utilizados

	Sólido Dissolvido Total (mg/L)	Alcalinidade Total (mg/L)	Oxigênio Dissolvido (mg/L)	Turbidez (NTU)
dados	5.649,0	5.649,0	5.649,0	5.649,0
média	1.120,7	46,2063	7,3041	21,213
desvio padrão	4.180,9	43,9598	1,8595	121,38
mín	0,0	1,0000	0,0	0,0
25%	56,000	20,0000	6,2600	3,3400
50%	1.000,0	33,0000	7,4300	7,7000
75%	160,00	60,0000	8,4400	17,000
máx	83.000	992,1000	14,270	67.770

Tabela 2 - Resumo estatístico dos parâmetros utilizados

	Nitrogênio Kjeldahl (mg/L)	Clorofila-a(µg/L)	Fósforo Total (mg/L)	Temperatura da Água (ºC)	pH
dados	5.649,0	5.649,0	5.649,0	5.649,0	5.649,0
média	0,6572	5,0449	0,092100	23,6614	7,2700
desvio-padrão	0,8278	5,5356	0,22260	5,1395	0,6321
mín	0,0160	0,010000	0,0020000	3,7500	4,5000
25%	0,3500	1,0700	0,0200	20,000	6,8500
50%	0,5000	2,6700	0,042000	24,000	7,2300
75%	0,7000	7,4800	0,082000	27,680	7,6600
máx	15,200	383,00	5,0000	36,200	9,9400

Em relação à concentração de fósforo total, o maior valor foi encontrado na bacia do rio Tennesse (ADEM, 2009) devido ao cultivo na agrícola utilizando agrotóxicos na área do ponto em que mostrou esse valor acima do permitido na maioria das legislações que é de cerca de 0,1 mg/L, com a CONAMA 357/05. As altas concentrações de fósforo total em água podem levar ao aumento de algas no corpo hídrico, diminuindo a concentração de oxigênio dissolvido para macroinvertebrados e peixes. Essa elevação dos nutrientes pode ocasionar o fenômeno de eutrofização das águas, conforme já comentado.

A variável clorofila que foi o alvo de predição deste estudo obteve um desvio padrão significativo já que os valores mínimos e máximos oscilaram entre 0,01 e 383 µg/L. Valores entre 2,5 e 10 µg/L são consideradas para a OMS em águas para consumo humano com tratamento mínimo, classe II. Em corpos hídricos como o Reservatório Taiuçupeba, região de Suzano – SP, com mais de 9 µg/L no ano de 2017, aponta para a eutrofização do reservatório devido a ocupação desordenada na região, ao uso intenso de defensores agrícolas e agrotóxicos que acabam escoando em direção ao reservatório, conforme já relatavam Arruda *et al.* (2014).

Assim como o reservatório Taiuçupeba, o reservatório Nances, no Alabama, sofre com a eutrofização devido ao uso da terra para a agricultura que acaba escoando parte dos fertilizantes utilizado para suas águas, bem como as atividades urbanas da cidade de Piedmont conforme publicação da *Geological Survey of Alabama* (2006), além do uso da terra naquela região para as pastagens, que podem ocasionar a eutrofização daquele corpo hídrico.

Os valores pH ficaram próximo ao valor considerado adequado para os corpos hídricos que é de 7. Valores acima de 7 foram encontrados tanto nos dados da CETESB quanto da USGS. No caso da CETESB, destaca-se o rio Paraíba do Sul com um pH de aproximadamente 10, na Região de São José dos Campos. Esse elevado valor pode ser associado a concentração de clorofila-a que passa dos 20 µg/L em 2010, já que a atividade algal é capaz de alterar o pH do meio, bem como a atividade urbana com despejos de efluentes industriais contendo soda caústica que podem elevar a alcalinidade do meio. Outro exemplo disso, é o Rio Suwanee, na Flórida, um dos maiores desse Estado, que no ano de 2002, uma alcalinidade de 141 mg/L de $CaCO_3$ considerado elevado no

corpo hídrico, e consequentemente um pH alcalino, de cerca de 9,22, essa região possui atividades diversas, dentre elas industrial. Essa observação, também, pode ser feita para a barragem localizada no Lago West Point que fica entre os estados da Georgia e Alabama, no rio Chattahoochee, que é utilizado para recreação e abastecimento, que obteve um pH 9,18.

A alcalinidade total pode ser associada a formação rochosa do corpo hídrico, os despejos que este receba, as atividades fotossintéticas que ocorram no meio, bem como a presença de matéria orgânica. Esse último fator pode ser atribuído ao rio Capivari, no ponto de monitoramento localizado na cidade de Jundiaí, que sofre com cargas de efluentes domésticos e industriais, os quais elevam a alcalinidade do meio (CETESB, 2009; OLIVEIRA *et al.*., 2014). A matéria orgânica presente de forma elevada no rio corroborou para uma alcalinidade de 342 mg/L de $CaCO_3$ em junho de 2009. O reservatório Bartlett, formado pelo represamento do Rio Verde no Arizona, apresenta altos valores de carbono orgânico total devido ao tempo de retenção na água no reservatório, uso da mesma para irrigação e deposição devido a tempestades de areia contribuem para o aumento dos níveis de carbono orgânico dissolvido, propiciando o aumento da matéria orgânica e, consequentemente, da produção de carbonatos o que eleva a alcalinidade da água conforme comentam Westerhoff e Anning (2000) e Barry *et al.* (2016).

Oxigênio Dissolvido (OD) é um parâmetro essencial para a vida e seus valores devem ser superiores a 5 mg/L conforme recomendado pela OMS. Em concentrações abaixo de 3 mg/L, o corpo hídrico é considerado em estado de hipoxia o que acaba afetando a vida dos organismos que dependem deste para sobreviver e realizar suas atividades, limitando a existência de peixes neste ambiente. Efeitos da eutrofização podem reduzir a concentração de oxigênio devido as oxidações promovidas devido ao aumento da concentração de nitrogênio em suas diferentes formas, diminuindo a disponibilidade para os organismos dependem do mesmo.

Na região estuarina, próxima a baía de Mobile, no Alabama, foram observados valores próximos a anoxia (ausência de oxigênio) devido atividades da indústria petrolífera na região, ao descarte de esgoto doméstico e aos avanços agrícolas na região conforme publicação da Mobile Bay Modelling Report (2012), obrigando peixes a migrarem para outras regiões. Similar a este caso, com valores de concentração abaixo de 1 mg/L, tem-se o rio Jaguari, especificamente, no

ponto de monitoramento localizado na cidade de Bragança Paulista, nos anos de 2014 e 2017. Este vem recebendo despejos de esgoto por anos diminuindo seu poder de depuração devido a contaminação e com casos de mortandade de peixe. Além disso, as atividades industriais na região, bem como a retirada da mata ciliar para uso agrícola, corroboram para a diminuição de oxigênio dissolvido neste rio (KUNLASAK *et al.*., 2013; CETESB, 2014; RAMOS, 2015).

O nitrogênio em suas diferentes formas, também, contribui para a eutrofização das águas e mortandade de peixes, principalmente, na sua forma amoniacal. Boa parte do nitrogênio presente na água advém de atividades antrópicas externas ao corpo hídrico, como o uso de fertilizantes na agricultura, o despejo de esgoto *in natura* e dos efluentes industriais. Entretanto, cianobactérias podem fixar nitrogênio do ar, e crescer rapidamente, alterando a qualidade da água, conforme abordam Tundisi e Matsumara-Tundisi (2011). O rio Tietê, em São Paulo, conhecido por sua elevada contaminação, chegou a apresentar valores de 15 mg/L de nitrogênio total, no ano de 2002, na cidade de Mogi das Cruzes, valor muito acima da legislação do CONAMA 357/05, 2,18 mg/L, e das recomendações de tolerância internacional da WHO de até 10 mg/L. Parte disso, deve-se as atividades hortifrutigranjeiras, pastagens naturais e cultivadas nas zonas rurais dos municípios de Mogi das Cruzes. Nesta região, que fica próxima à cabeceira do rio, o nitrogênio utilizado na adubação e fertilizantes, acaba escoando durante a chuva provocando o aumento desse nutriente no corpo hídrico. (CETESB, 2002). O esgoto e o lixo urbano agravam a situação, conforme ilustrado na Figura 7.

Outro exemplo, da incapacidade de assimilação de um manancial a cargas elevadas de nutrientes advindo da agricultura e escoamento dos pastos, localiza-se no canal Ihagee, na bacia do rio Chattahoochee, no estado do Alabama, que vem sofrendo durante anos com tais atividades antrópicas (ADEM, 2006; ADEM, 2014).

Figura 7- Proliferação de algas resultante da concentração de carga orgânica proveniente do esgoto não tratado e despejado no manancial
Fonte: O Diário de Mogi (2018)

A turbidez é um dos principais parâmetros na análise da qualidade da água. A presença de algas acaba afetando a capacidade de luz na água, interferindo nos organismos que necessitam da mesma para sobreviver. Sólidos em água devido a erosão, matéria orgânica e poluentes diversos contribuem para a variação da turbidez que possui um valor desejado para corpos hídricos que necessitam de tratamento de água de até 40 NTU conforme legislação brasileira, e 5 NTU pela WHO. Elevados valores de turbidez, portanto, implicam em maiores gastos na adição de cloro, por exemplo, nas estações de tratamento, encarecendo o valor da mesma ao ser distribuída.

Na base de dados formada, foram encontrados valores muito acima dos que as legislações permitem, como o caso do Rio Una, na captação da SABESP, em Taubaté -SP, com mais de 3000 NTU. Consta no relatório da CETESB, do ano de 2012, o rio apresentou elevados valores de metais com Alumínio, Ferro, Cromo e Níquel, além, de receber uma elevada carga de esgoto das cidades circunvizinhas. Além de todos esses fatores, o rio ainda sofre com o assoreamento natural em suas margens (GONÇALVES, 2016). O rio Pardo, na captação da SABESP, em Ourinhos-SP, também,

apresentou uma elevada turbidez de mais de 2000 NTU, devido ao grande despejo de esgotos, uso excessivo do solo, com consequente escoamento de fertilizantes da parte superficial do solo, aumentando a eutrofização do corpo hídrico, sendo este, também, motivo da elevada turbidez, conforme relatórios da CETESB em 2012 e 2015. Com mais de 150 NTU, o rio Pea, no estado do Alabama, próximo à fronteira com a Flórida, sofre muito com a sedimentação nas suas margens, bem como a atividade agrícola, o que eleva consideravelmente os níveis de turbidez neste corpo hídrico, conforme relatam Murgulet e Cook (2010). Ainda no estado do Alabama, próxima a cidade de Birmingham, o rio Cahaba, sofre com excesso de nutrientes devido as atividades agrícolas, aparecimento de algas diminuindo a zona fótica do corpo hídrico, elevando a turbidez do meio, registro do ano de 2006.

3.3.2 Redes Neurais

Os dados avaliados anteriormente serviram de base para a utilização de redes neurais com séries temporais, para a previsão de clorofila-a. Na análise das redes neurais utilizadas foram observados que os valores de treino para o *dataset* utilizado, obtiveram boas métricas de predição, conforme apresentado na Figura 8.

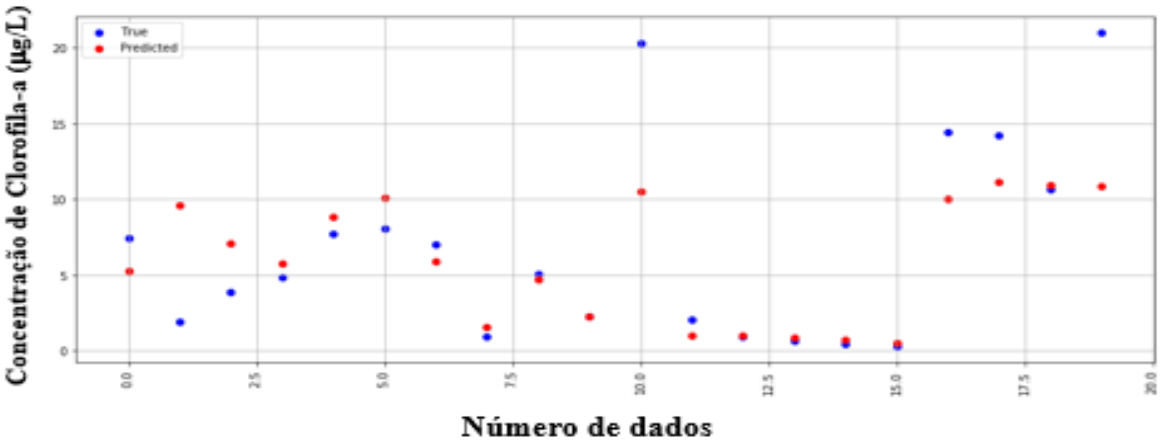


Figura 8- Dados de treinamento – Concentração de Clorofila-a (µg/L)

Já para os valores de teste utilizado, também, obteve-se valores próximos dos mensurados em campo, mostrando a precisão da rede, conforme Figura 9.

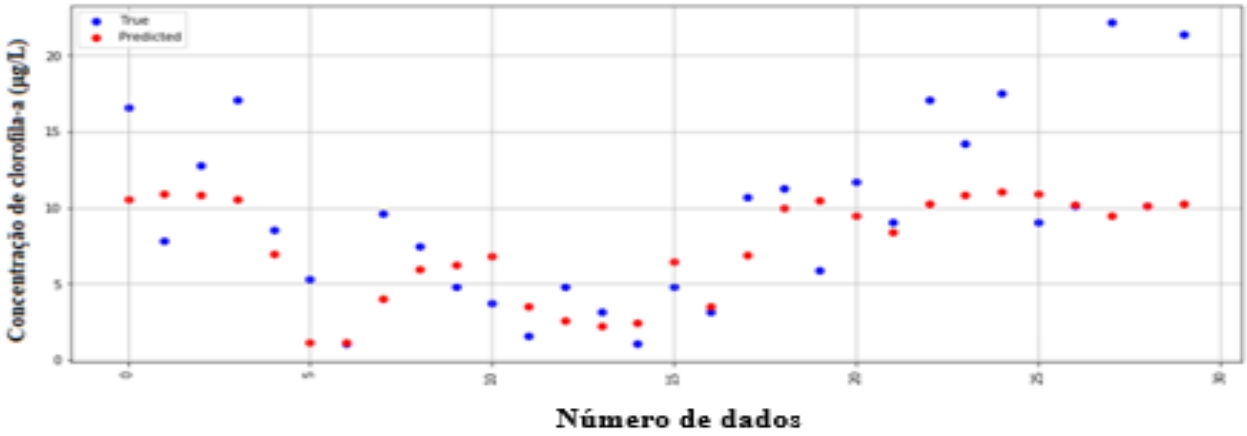


Figura 9 - Dados de teste - Concentração de Clorofila-a (µg/L)

Após a verificação da distribuição dos valores preditos e reais, pode-se calcular as métricas de MAE e RMSE. Quanto mais próximo de zero maior a precisão. O valor de MAE foi igual a 0,3731 e o do RMSE 0,5118. Para que se tenha uma noção ainda maior do que representa essas métricas, alguns dos valores reais e preditos são apresentados na Tabela 3 para que essa diferença seja ainda mais sensível. Park *et al.* (2015) encontraram valores de MAE igual a 0,52 e 5,72 no forecasting de clorofila-a avaliados em reservatórios da Coréia do Sul, mostrando que a rede neural desenvolvida no presente trabalho apresenta resultados melhores do que alguns encontrados na literatura.

Rajaee e Boroumand (2015) realizaram trabalho para previsão (*forecasting*) clorofila-a utilizando apenas as séries temporais de clorofila-a e encontraram um RMSE de 1,58. Estes autores sugeriram trabalho similar ao desenvolvido na presente dissertação, ou seja, utilizar séries temporais de mais variáveis ambientais para que o *forecasting* observasse as relações da concentração de clorofila-a com as outras variáveis e os impactos que estas poderiam trazer aos resultados.

Tabela 3- Valores reais e preditos da concentração de clorofila-a (µg/L)

Valor Real (µg/L)	10,1	7,48	8,54	10,1	7,12	7,48	13,4	4,63	3,47	6,41	11,7
Valor Predito (µg/L)	10,56	8,35	10,52	10,78	10,51	10,43	9,57	5,82	1,28	6,25	9,95

Neste trabalho, como já citado anteriormente, foram utilizados três atrasos para o *forecasting* de clorofila-a, o que significa que o modelo foi construído para predizer a concentração

de clorofila-a em uma campanha conseguinte (t+1) com base em uma campanha atual e nas três anteriores. Observa-se que mesmo como um *dataset* sendo reduzido devido alguns outliers, a predição ainda assim se tornou precisa.

Pela análise de correlação de Pearson, Figuras 10 (a, b e c) e 11, observa-se uma maior correlação da clorofila-a com o pH, temperatura, nitrogênio total, alcalinidade total e fósforo total. Os menores valores foram observados para o oxigênio dissolvido, turbidez e inversa com sólidos totais dissolvidos. A maior correlação com pH com a clorofila-a deve-se a fotossíntese realizada pelos fitoplânctons e algas nos corpos hídricos o que faz alterar a alcalinidade do meio (TALLING, 2010; PHILLIPS *et al.*, 2015). Essa última característica, por consequência, também teve alta correlação com a clorofila-a. Outra correlação importante foi com o nitrogênio total em que há a relação de antecedente e consequente, nitrogênio e clorofila-a, respectivamente. O nitrogênio está associado ao escoamento superficial de fertilizantes relacionados as atividades agrícolas que são desenvolvidas em áreas próximas aos corpos hídricos e que nos períodos chuvosos acabam sendo levados aos mananciais pela lixiviação, conforme citam Srebotnjak *et al.* (2012) e Rajaee e Boroumand (2015).

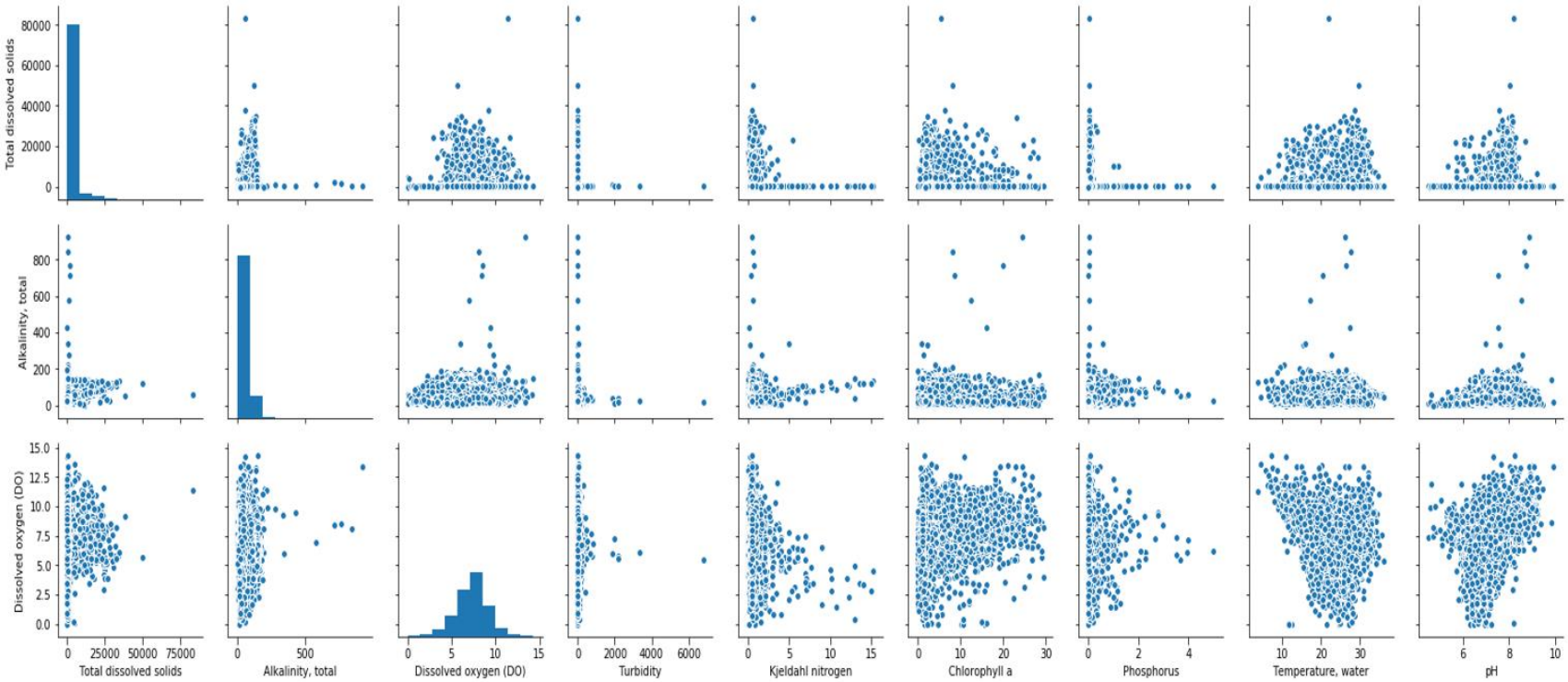


Figura 10 a- Correlação de Pearson entre as variáveis analisadas

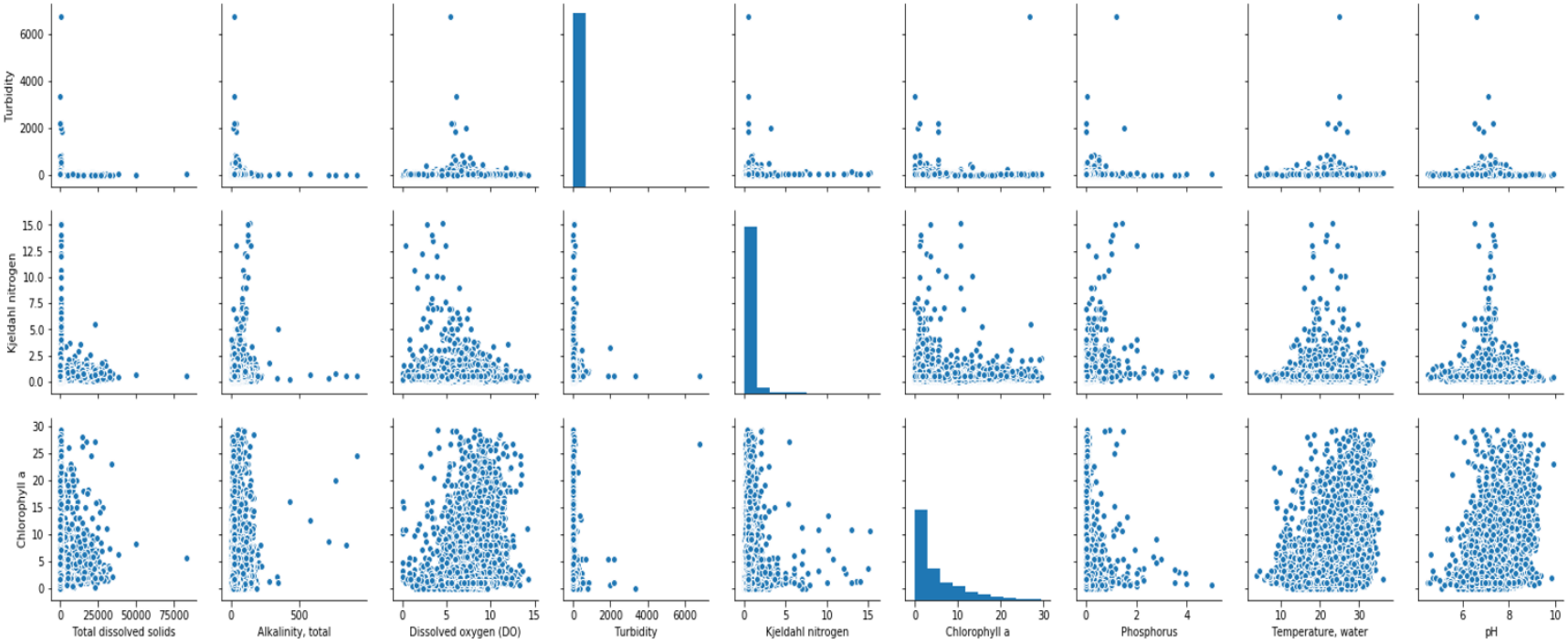


Figura 11 b- Correlação de Pearson entre as variáveis analisadas

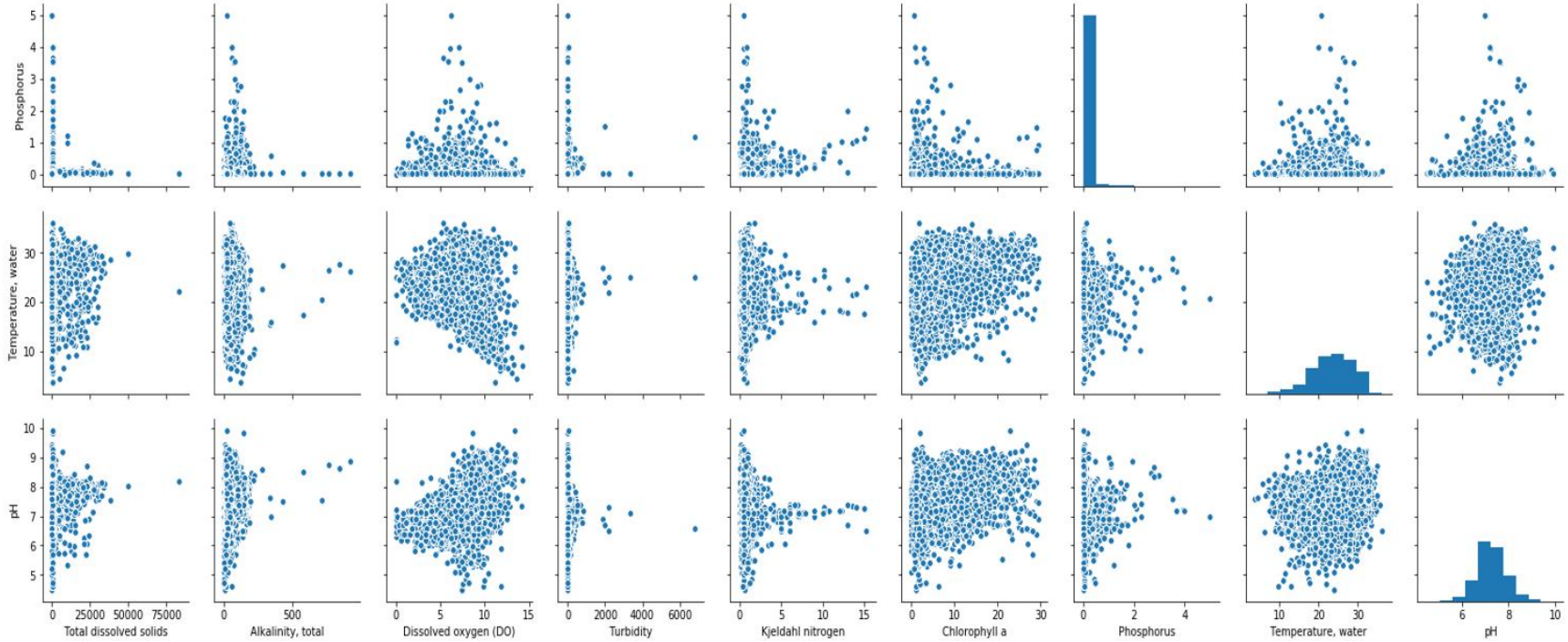


Figura 12 c- Correlação de Pearson entre as variáveis analisadas

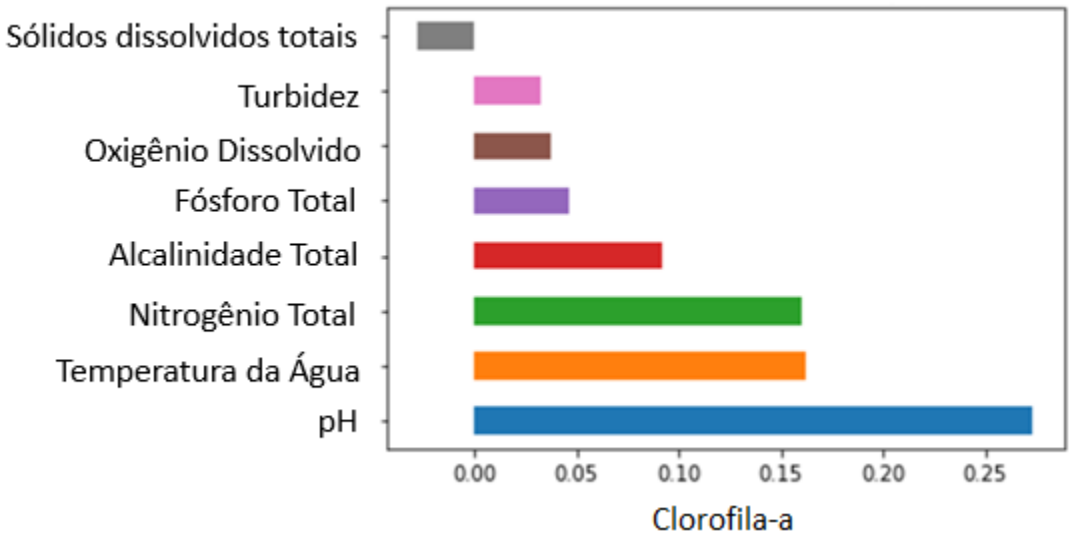


Figura 13– Score de correlação entre as variáveis de previsão com a clorofila-a

O *forecasting* realizado de acordo com os resultados das redes mostrou-se bastante preciso e consistente com os fenômenos físico-químicos. Observa-se que o monitoramento da CETESB, embora, recente, em relação ao da USGS-EPA, já mostra bastante efetivo quanto aos resultados e a disponibilidade dos dados. Embora ocorra falta de dados ou dados mal registrados, ambos monitoramentos servem de exemplo para as agências em todo mundo. Além disso, observa-se que as características de descarte e uso da terra, diferenciando apenas as estações mais rigorosas e definidas no sul dos EUA que no Brasil. No entanto, isso não influenciou o modelo quanto aos resultados, pois, as influências das variáveis são semelhantes em ambos os locais.

O objetivo de elaboração de um modelo de *forecasting* que pode ser utilizado em diversas partes do mundo foi alcançado neste trabalho, e o mesmo possibilita a adição de novos dados para que a sua base aumente e, por consequência, a precisão do mesmo. Este modelo não é o solucionador dos problemas relacionados a eutrofização, mas uma proposta de caminho que pode ser tomado para representar o comportamento fenomenológico de um corpo hídrico e assim poder elaborar estratégias que possam reduzir custos com material laboratorial e de mão de obra.

O fenômeno de eutrofização, por ser um dos fenômenos que mais preocupa as agências de qualidade da água, influenciou na escolha da clorofila-a, como variável *target* deste trabalho. Este

fenômeno afeta não apenas o tratamento da água e o abastecimento, mas até mesmo corpos hídricos que são usados somente para harmonia paisagística e navegação, já que mesmo estes sendo destinados a essas atividades, há o contato direto das pessoas o que pode ser uma via de transmissão de doenças, fato este enfatizado por Budria *et al.* (2017).

A ampliação do monitoramento e alertas para o uso da terra e descarte poderia ser mais efetivos nas estações analisadas deste trabalho, em que corpos hídricos do Alabama e das zonas industriais de São Paulo, mostraram uma clara necessidade de maior cuidado por parte dos usuários e do governo. Esses corpos hídricos sofrem com cargas de nitrogênio, principalmente, que pode acarretar em uma eutrofização sem volta, e há projeções de que a região sul da bacia do Rio Mississipi, um dos principais dos EUA, sofra com um nível excessivo de nitrogênio devido ao aumento das chuvas, principalmente, devido as cargas de seus afluentes, conforme ressalta Sinha *et al.* (2017).

Sendo assim, o monitoramento da qualidade das águas é um dos fatores que pode contribuir para a geração de informação para a população a respeito da qualidade das águas. Os registros desse monitoramento possibilitam a análise e elaboração de relatórios que servem de alerta para os atores envolvidos em uma bacia hidrográfica. Para isso, faz-se necessária uma maior padronização das análises, e principalmente, dos registros para que usuários secundários possam utilizar para o benefício da sociedade. Além disso, o compartilhamento dos dados deve ser mais frequente com menos necessidade de atender interesses específicos, o que pode auxiliar na minimização de erros no registro e comparação entre as bases de dados para modelos de *machine learning* como o do presente trabalho sejam desenvolvidos.

3.3.3 Estudo de caso: Bacia Paraíba do Sul – Ponto Jacareí

Para colaborar com a estratégia adotada, os mesmos parâmetros utilizados para a predição de clorofila-a da base de dados USGS/CETESB, foi possível obter séries temporais para uma bacia específica monitorada pela CETESB, como exemplo a bacia do Rio Paraíba do Sul, no município de Jacareí (PARB02200), e construir uma rede neural nos mesmos moldes da anterior. Os valores encontrados de MAE ficam bem próximos de zero o que confirma a aproximação dos valores reais

e os preditos, mostrando que apesar da quantidade de dados com os parâmetros utilizados especificamente não ser elevada, a rede conseguiu mostrar bons resultados.

Segundo dados do relatório de Águas Interiores do Estado de São Paulo CETESB (2017) a qualidade da água na bacia encontra-se com Índice de Qualidade de Água (IQA) classificado como boa na média dos anos de 2012-2016 no munícipio de Jacareí, mostrando que apesar de ações antrópicas na região a qualidade da água não foi muito alterada. No que diz respeito aos níveis de clorofila-a, o período de monitoramento (2008-2017) obteve uma média de 0,55 µg/L, abaixo dos 10 µg/L para corpos hídricos da Classe 1 conforme a resolução 357/2005 do CONAMA, mostrando que apesar das atividades antrópicas na região, não houve uma variação muito grande das concentrações deste parâmetro. A Tabela 4 apresenta a estatística descritiva dos dados analisados.

Tabela 4 – Estatística descritiva dos parâmetros ambientais no ponto Jacareí

	Sólido Dissolvido Total (mg/L)	Alcalinidade Total (mg/L)	Oxigênio Dissolvido (mg/L)	Turbidez (NTU)	Nitrogênio Kjeldahl (mg/L)	Clorofila-a (µg/L)	Fósforo Total (mg/L)	Temperatura da Água (ºC)	pH
dados	40,0	40,0	40,0	40,0	40,0	40,0	40,0	40,0	40,0
média	61,4	2,23	6,24	19,9	0,560	0,590	0,0500	22,7	6,92
desvio padrão	22,1	378	1,07	16,1	0,150	0,540	0,0300	3,01	0,290
mín	50,0	2,00	3,40	5,20	0,500	0,0100	0,0100	18,0	6,20
25%	50,0	2,00	5,60	11,0	0.5	0,0100	0,0300	20,0	6,80
50%	50,0	2,00	6,40	15,0	0,500	0,530	0,0400	22,6	6,90
75%	61,5	2,30	6,85	23,2	0,50	1,00	0,07	25,8	7,10
máx	144	3,43	8,80	96,0	1,16	1,50	0,150	28,0	7,74

Por meio da correlação de Pearson, pode-se observar a influência da alcalinidade na concentração de clorofila-a neste ponto, bem como dos sólidos dissolvidos totais (SDT) e o fósforo total conforme abordado na Figura 12. Ao mesmo tempo, observa-se uma relação inversa entre as concentrações de Oxigênio Dissolvido (OD) e de clorofila-a, ou seja, quanto maior a concentração de clorofila-a menor será a de OD, o que demonstra que a presença de elevadas concentrações de

clorofila-a implicam na existência de algas e essas para realizarem a fotossíntese necessitam de consumo do oxigênio diminuído sua concentração.

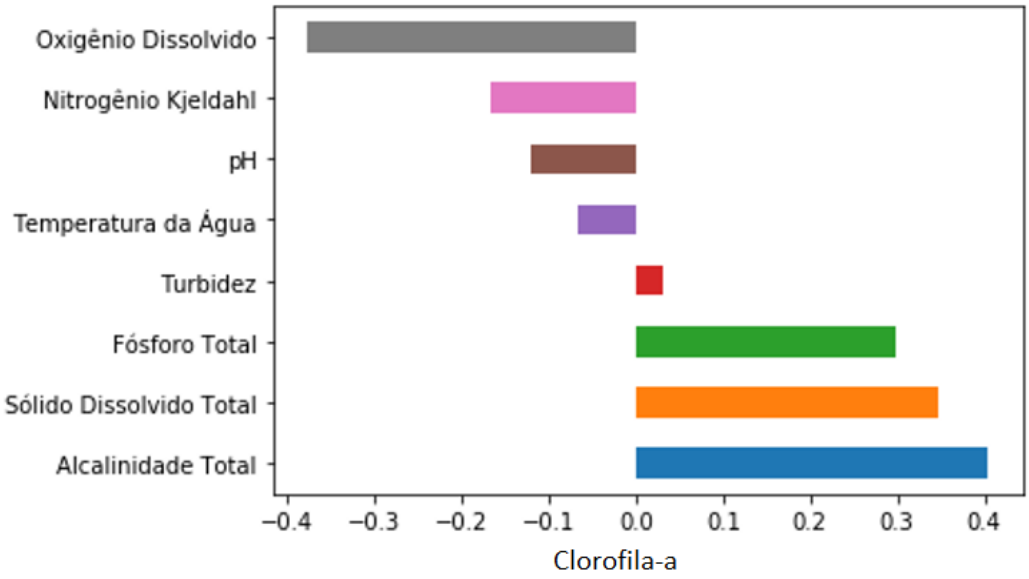


Figura 14- Score de correlação entre as variáveis de previsão com a clorofila-a

Para a alcalinidade total essa relação é explicada pela existência de carbonatos e hidróxidos que podem aparecer em águas nas quais ocorrem florações de algas (eutrofizadas), sendo que em período de intensa insolação o saldo da fotossíntese em relação à respiração é grande e a retirada de gás carbônico provoca elevação de pH para valores altos. A principal fonte de alcalinidade de hidróxidos em águas naturais decorre da descarga de efluentes de indústrias, como na região de Jacareí, que ainda sofre influência da chegada das águas das regiões anteriores do Rio Paraíba em pode se empregam bases fortes como soda cáustica e cal hidratada. Em águas tratadas, pode-se registrar a presença de alcalinidade de hidróxidos em águas abrandadas pela cal (CETESB, 2017).

O resultado da correlação de Pearson mostra que as concentrações de fósforo estão relacionadas diretamente com a clorofila-a, já o inverso é observado para a concentração de nitrogênio Kjedahl, devido as baixas concentrações de clorofila-a e fósforo total, conforme observado na Tabela 4, e a de nitrogênio ser elevada. Os valores de nitrogênio podem apresentar este valor elevado devido a diversas fontes de contaminação, como descarte de esgotos domésticos

da região (efluentes que possuam nitrogênio em sua composição) e deposição atmosférica devido as atividades antrópicas ao redor da bacia. As concentrações de nitrogênio, inclusive, estão acima dos valores exigidos pela resolução CONAMA 357/2005.

Similar ao que foi feito no tópico anterior para uma base de dados maior, foi realizada a alimentação da rede com dados dos parâmetros e suas séries temporais. Conforme apresentado na metodologia, a rede foi treinada baseada na série temporal dos dados do ponto Jacareí na bacia do rio Paraíba do Sul. A Figura 13 apresenta um gráfico com pontos de treinamento, que foram 80% da base de dados utilizada. Os pontos em azul representam os valores mensurados e os pontos em vermelhos, os pontos preditos pela rede. Os valores absolutos para essa etapa estão apresentados na Tabela 5.

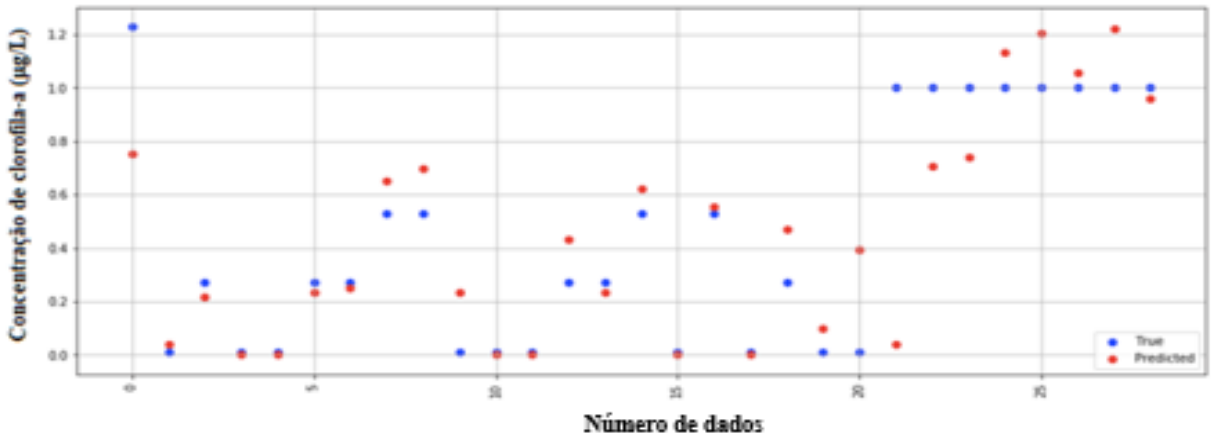


Figura 15 – Dados de Treinamento – concentração de clorofila-a no ponto Jacareí

Tabela 5- Valores de treinamento da rede – concentração de clorofila-a no ponto Jacareí

Valor Real (µg/L)	0,01	0,27	0,27	0,27	0,53	0,27	0,53	0,53	0,27	0,01
Valor Predito (µg/L)	0,04	0,21	0,23	0,25	0,66	0,23	0,62	0,55	0,47	0,099

Em relação aos 20% dos dados analisados restantes, que serviram de teste para o modelo, observou-se uma aproximação significativa desses 8 pontos, conforme ilustra a Figura 14. A

Tabela 6 apresenta os valores de concentração mensurados para clorofila-a e os valores preditos da rede.

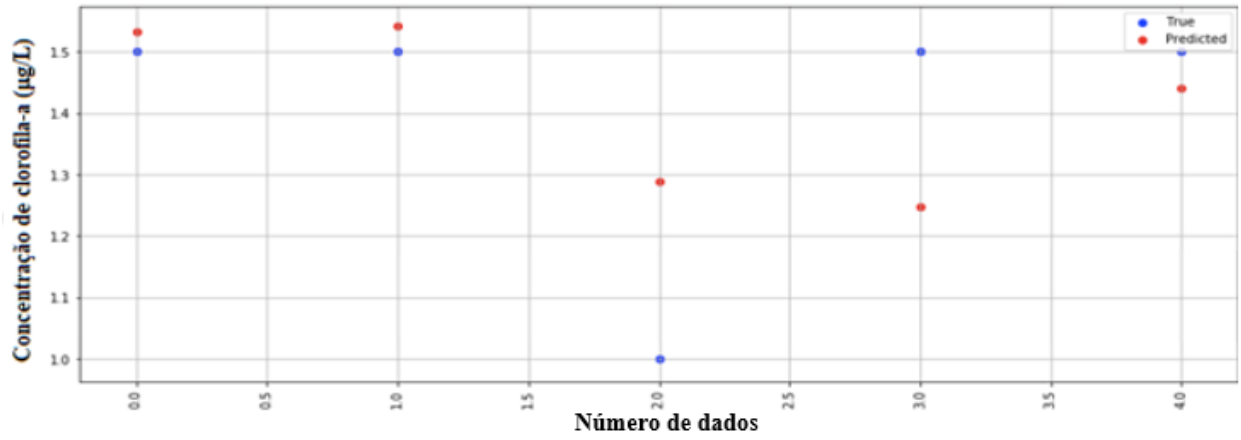


Figura 16- Rede Neurais Artificiais - Teste – ponto Jacareí

É importante comentar ainda sobre os valores de teste da rede para o ponto Jacareí que os valores reais da concentração de clorofila-a praticamente iguais não significa que a rede está sobreajustada, pois os valores preditos validam a convergência da rede.

Tabela 6- - Valores de teste da rede – concentração de clorofila-a no ponto Jacareí

Valor Real (µg/L)	1,5000	1,5000	1,0000	1,5000	1,5000
Valor Predito (µg/L)	1,5331	1,5414	1,2887	1,2476	1,4398

O Erro Médio Absoluto (MAE) foi de 0,3974 e a Raiz do Erro Quadrático Médio (RMSE) igual a 0,5313, mostrando bastante precisão para os dados apresentados. Levando em consideração que apenas um ponto de um rio originou essa análise robusta, esses valores de erro estão dentro do aceitável, principalmente, o MAE que indica a média das diferenças do valor real e do valor predito pela rede.

Conforme citado por Srebotnjak *et al.* (2012), os métodos matemáticos auxiliam a compreender os fenômenos e as tomadas de decisão, mas para isso é necessário um aumento e melhoria do monitoramento como forma crucial de garantir a qualidade da água, abastecimento

para população e redução das doenças que estejam relacionadas ao consumo de água. Sendo assim, este estudo traz esse alerta para as autoridades e os órgãos responsáveis pela a análise e abastecimento de água.

Em um contexto geral, Tundisi e Matsumara-Tundisi (2011) ressaltam que um corpo hídrico em baixo nível de eutrofização pode ser utilizado até certo ponto para a geração de energia elétrica ou irrigação, mas para abastecimento público ou recreação exige-se tratamento prévio. Sendo assim, um corpo hídrico por sua eutrofização, o que se deve fazer é buscar formas de mitigar os impactos os quais o mesmo é submetido, diminui-los ou prever quando isso pode acontecer, para que os múltiplos usos sejam atendidos.

Além disso, o método mostrou-se um ótimo caminho para a redução de análises laboratoriais, já que a análise de clorofila-a por ser uma das mais custosa reduziria significativamente o orçamento para análises laboratoriais, dando oportunidade para outro destino a esse passivo. A utilização de redes neurais artificiais é um caminho para entendimento e predição de dados ambientais, carregando consigo precisão e facilitando a compreensão dos gestores sobre os diversos fenômenos que ocorrem em uma bacia hidrográfica. Esse estudo possibilitou rápido entendimento das relações entre os parâmetros ambientais e os eventos que influenciam suas concentrações ao longo dos anos.

3.4 CONCLUSÃO

A aplicabilidade de modelos de *machine learning* para auxiliar a gestão dos recursos hídricos é algo inegável ao redor do mundo. Nesse sentido, este trabalho apresentou um modelo de redes neurais com alimentação exógena (NARX) utilizando séries temporais de variáveis de qualidade da água e usando a concentração de clorofila-a, *target*, para descrever o comportamento ambiental de diferentes corpos hídricos. O trabalho utilizou a base de dados de três estados do sul dos EUA, disponíveis no *Water Quality Portal* pela USGS/EPA, e dados disponibilizados pela CETESB. Esses dados foram filtrados e os parâmetros que mais foram medidos, juntamente com a clorofila-a, foram alcalinidade total, fósforo, nitrogênio Kjedahl, sólidos dissolvidos totais, pH, temperatura, turbidez e oxigênio dissolvido, resultando em um conjunto de dados para aplicação da NARX.

A NARX apresentou, então, resultados de concentração de clorofila próximos aos valores mensurados em campo pelas agências de cada estado dos EUA e do estado de São Paulo, indicando o excelente desempenho da rede comprovado pelos baixos valores de MAE e RMSE, parâmetros estes que serviram como análise de sensibilidade do modelo. A partir desses resultados de desempenho, aplicou-se a rede a um caso específico da CETESB, um ponto no rio Jacareí, na bacia do Rio Paraíba do Sul, para validação da rede. Os resultados dessa validação corroboraram com valores de MAE e RMSE indicando que a predição de clorofila-a através da rede foi satisfatória.

Para que modelos como este sejam elaborados e mais precisos, com uma visão ainda mais global, é necessária uma uniformização nos registros dos parâmetros de qualidade da água para facilitar a análise e a proposta de modelos. Essa não uniformização foi uma das dificuldades encontradas na elaboração deste trabalho.

Em conformidade com a proposta deste trabalho, vale salientar que o modelo pode incorporar dados de outras agências do mundo, bem como de outras variáveis, para que este seja ainda mais abrangente e sua aplicação para descrever o comportamento dos corpos hídricos seja maior. Este modelo pode contribuir não somente para a previsão de variáveis e redução de custos laboratoriais, mas também como um benefício social de melhoria da qualidade da água. Sugere-se que em trabalhos futuros outros modelos de redes neurais sejam treinados com previsão de uma outra variável que não seja a clorofila-a para que o modelo possa atender os interesses de outros atores no uso da água.

REFERÊNCIAS

ADEM, Alabama Department of Environmental Management, 2014. Disponível em: http://adem.alabama.gov/programs/water/wqsurvey/table/2014/2014IhageeCk_IHGR-1.pdf . Acesso em: 01/10/2018.

ADEM, Alabama Department of Environmental Management. 2009 Disponível em: http://adem.alabama.gov/programs/water/wqsurvey/table/2009/2009SwanCk.pdf. Acesso em: 01/10/2018.

ADEM, Alabama Department of Environmental Management. Surface Water Quality Screening Assessment of the Southeast Alabama River Basins- 2006

ARRUDA, R.D.O.M., dos SANTOS, M.A., VIPPER, H.P.A.F. and de Souza ROCHA, M. AVALIAÇÃO DA QUALIDADE DO RESERVATÓRIO DE TAIAÇUPEBA, MOGI DAS CRUZES-SP, SOB O ASPECTO DA SAÚDE, ENTRE 2009 E 2013. **Revista Geociências-UNG-Ser**, *13*(1), pp.38-49, 2014.

BARRY, M., CHIU, C.A.; WESTERHOFF, P.. Severe weather effects on water quality in Central Arizona. **Journal-American Water Works Association**, *108*(4), pp.E221-E231, 2016.

BUDRIA, A. Beyond troubled waters: the influence of eutrophication on host–parasite interactions. **Functional Ecology**, v. 31, n. 7, p. 1348-1358, 2017.

CETESB, Companhia Ambiental do Estado de São Paulo. Disponível em https://cetesb.sp.gov.br/aguas-interiores/publicacoes-e-relatorios/>. Acesso em: 15/07/2018.

CETESB, Companhia Ambiental do Estado de São Paulo. Disponível em https://cetesb.sp.gov.br/aguas-interiores/publicacoes-e-relatorios/>. Acesso em: 15/07/2018.

CETESB. Relatório de Qualidade das águas interiores no estado de São Paulo, (Série Relatórios / CETESB, ISSN 0103-4103), 2017.

CETESB. Relatório de Qualidade das águas interiores no estado de São Paulo, (Série Relatórios / CETESB, ISSN 0103-4103), 2012.

CETESB. Relatório de Qualidade das águas interiores no estado de São Paulo, (Série Relatórios / CETESB, ISSN 0103-4103), 2014.

CETESB. Relatório de Qualidade das águas interiores no estado de São Paulo, (Série Relatórios /CETESB, ISSN 0103-4103), 2015.

CETESB. Relatório de Qualidade das águas interiores no estado de São Paulo, (Série Relatórios /CETESB, ISSN 0103-4103), 2009.

CHANG, F. J.; TSAI, Y. H.; CHEN, P. A.; COYNEL, A.; VACHAUD, G. Modeling water quality in an urban river using hydrological factors–Data driven approaches. **Journal of environmental management**, *151*, 87-96,2015

CHOU, Jui-Sheng; HO, Chia-Chun; HOANG, Ha-Son. Determining quality of water in reservoir using machine learning. **Ecological Informatics**, v. 44, p. 57-75, 2018.

COOK, M.R.; MOSS, N.E. Analysis of Discharge and Sediment Loading Rates in Tributaries of Dog River in the Mobile Metropolitan Area. Open File Report. Tuscaloosa, Alabama. 2012

COSTA, J. A. D.; SOUZA, J. P. D.; TEIXEIRA, A. P.; NABOUT, J. C.; CARNEIRO, F. M. Eutrophication in aquatic ecosystems: a scientometric study. **Acta Limnologica Brasiliensia**, *30*, 2018.

GEOLOGICAL SURVEY OF ALABAMA, Disponível em: https://www.gsa.state.al.us/img/Ecosystems/pdf/OFR_0601.pdf. Acesso em: 01/10/2018.

GONÇALVES, F.M. Bacia Hidrográfica do Rio Paraíba do Sul: Avaliação Integrada da Qualidade das Águas dos Estados de Minas Gerais, Rio de Janeiro e São Paulo. Dissertação de Mestrado para o programa Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais, 2016.

KHAN, Y; SEE, C. S. Predicting and analyzing water quality using Machine Learning: A comprehensive model. In: **Systems, Applications and Technology Conference (LISAT), 2016 IEEE Long Island**. IEEE, 2016. p. 1-6.

KUNLASAK, K., CHITMANAT, C., WHANGCHAI, N., PROMYA, J. AND LEBEL, L. Relationships of dissolved oxygen with chlorophyll-a and phytoplankton composition in tilapia ponds. **International Journal of Geosciences**, *4*(05), p.46, 2013.

MOBILE BAY MODELLING REPORT, 2012, Disponível em: http://www.mobilebaynep.com/images/uploads/library/Mobile_Bay_Modeling_Report_(2012).pdf Acesso em: 02/10/2018

MONTEIRO, M.; COSTA, M. A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. **Hydrology**, v. 5, n. 3, p. 1-20, 2018.

MONTEIRO, M.; COSTA, M. A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. **Hydrology**, v. 5, n. 3, p. 1-20, 2018.

MURGULET, D.; COOK, M.R.Water-Quality Evaluation of the Choctawhatchee and Pea Rivers in Southeast Alabama – **Geological Survey of Alabama**, 2010.

NODOUSHAN, E.J. Monthly Forecasting of Water Quality Parameters within Bayesian Networks: A Case Study of Honolulu, Pacific Ocean. **Civil Engineering Journal**, v. 4, n. 1, p. 188-199, 2018.

NODOUSHAN, E.J. Monthly Forecasting of Water Quality Parameters within Bayesian Networks: A Case Study of Honolulu, Pacific Ocean. **Civil Engineering Journal**, v. 4, n. 1, p. 188-199, 2018.

O Diário de Mogi, 2018. Disponível em: https://www.odiariodemogi.net.br/rio-tiete-desaparece-sob-tapete-de-plantas-aquaticas-em-mogi/. Acesso em: 02/10/2018

OLIVEIRA, H., MORTATTI, J., DE MORAES, G. M., VENDRAMINI, D., & DE GENOVA CAMPOS, K. B. Caracterização físico-química da carga dissolvida dos rios Jundiaí e Capivari, São Paulo. **Geochimica Brasiliensis,** *28*(1), 23-35, 2014.

PARK, Y.; CHO, K. H.; PARK, J.; CHA, S. M.; KIM, J. H. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. **Science of the Total Environment**, *502*, 31-41, 2015.

PHILLIPS, J. C.; MCKINLEY, G. A.; BENNINGTON, V.; BOOTSMA, H. A.; PILCHER, D. J.; STERNER, R. W.; URBAN, N. R. The potential for CO2-induced acidification in freshwater: A great lakes case study. **Oceanography**, v. 28, n. 2, p. 136-145, 2015

RAJAEE, T; BOROUMAND, A. Forecasting of chlorophyll-a concentrations in South San Francisco Bay using five different models. **Applied Ocean Research**, v. 53, p. 208-217, 2015.

RAMOS, M.A.G. Avaliação da qualidade da água dos rios Jaguarí e Atibaia por meio do índice de qualidade da água-IQA e ensaios toxicológicos, Tese de doutorado – Instituto de Biociências (UNESP), 2015.

RICKWOOD, C.J.; CARR, G.M. Development and sensitivity analysis of a global drinking water quality index**. Environmental monitoring and assessment**, *156*(1-4), p.73, 2009.

SINHA, E.; MICHALAK, A. M.; BALAJI, V. Eutrophication will increase during the 21st century as a result of precipitation changes. **Science**, v. 357, n. 6349, p. 405-408, 2017.

SREBOTNJAK, T.; CARR, G.; DE SHERBININ, A.; RICKWOOD, C. A global Water Quality Index and hot-deck imputation of missing data. **Ecological Indicators**, *17*, pp.108-119, 2012.

USGS, United States Geologcial Service. Disponível em <https://www.waterqualitydata.us/>. Acesso em: 20/07/2018.

USGS, United States Geological Service. Disponível em <https://www.waterqualitydata.us/>. Acesso em: 20/07/2018.

VERMA, A.K.; SINGH, T.N. Prediction of water quality from simple field parameters. **Environmental earth sciences**, *69*(3), pp.821-829, 2013.

WALSH, K. Dog river water quality: before and after precipitation. Disponível em: http://www.usouthal.edu/geography/fearn/480page/2011/11Walsh.pdf. Acessado em : 26/09/18.

WESTERHOFF, P.; ANNING, D. Concentrations and characteristics of organic carbon in surface water in Arizona: influence of urbanization. **Journal of hydrology**, *236*(3-4), pp.202-222, 2000.

TALLING, J. F. pH, the CO_2 system and freshwater Science. **Freshwater Reviews**, v. 3, n. 2, p. 133-147, 2010.

TUNDISI, J.G.; MATSUMURA-TUNDISI, T. Recursos hídricos no século XXI. **Oficina de Textos**, 2011.

ZHANG, Y.; HUANG, J. J.; CHEN, L.; QI, L. Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China. **Journal of Hydroinformatics**, *17*(4), 679-695, 2015.

4 PREDIÇÃO DA CONCENTRAÇÃO DE CLOROFILA-A UTILIZANDO RANDOM FOREST

RESUMO

A avaliação da qualidade da água visa beneficiar e atender os usos múltiplos da água que desde muito tempo vem se degradando devido ao compasso acelerado de contaminação e mal uso da mesma. Os corpos lênticos, que servem para atender ao abastecimento humano, a irrigação entre outros, são as mais sofrem em termos de degradação ambiental. Visando facilitar o entendimento dos fenômenos de contaminação muitas variáveis físico, químicas e biológicas são analisadas e por não serem variáveis que se comportam de forma linear, necessitam de uso de um ferramental mais robusto para uma melhor compreensão. Sendo assim, este trabalho adotou o uso do *Random Forest* para a predição de clorofila-a para que fossem estabelecidas correlações entre as variáveis que algumas vezes a estatística tradicional não deixa claro. Este trabalho avaliou dados de reservatórios do estado de Sergipe buscando predizer a concentração de clorofila-a por meio de parâmetros ambientais, tais como, concentrações de fósforo e nitrogênio, DBO e OD. Na correlação de Spearman, a clorofila-a obteve maior correlação com o fósforo total e no Random Forest, através da análise de importância de cada variável (*feature importance*), o maior score foi para a DBO, o que sustentou a ideia de que este algoritmo quando aliado a estatística tradicional pode apresentar correlações não tão observadas na mesma. Os resultados apresentados foram aceitáveis segundo as métricas de análise de performance MAE (5,1914), RMSE (3,8473) e Oobscore (0,2288). Os valores apresentados poderiam ser ainda melhores se houvesse mais dados disponíveis, no entanto, o monitoramento na região ainda é recente.

Palavras-chave: Random Forest; Clorofila-a; Qualidade da água.

ABSTRACT

The assessment of water quality aims to benefit and meet the multiple uses of water that has long been degrading due to the accelerated pace of contamination and misuse of water. The lentic bodies, which serve to supply human necessities, irrigation among others, are the most suffering in terms of environmental degradation. In this sense, in order to facilitate the understanding of contamination phenomena many physical, chemical and biological variables are analyzed and because they are not variables that behave in a linear way, it is necessary to use of a more robust tool for a better understanding. Thus, this work adopted the use of Random Forest for the prediction of chlorophyll-a to establish correlations between variables that sometimes the traditional statistics does not make clear. So, this work evaluated data from reservoirs in the state of Sergipe, Brazil seeking to predict chlorophyll-a concentration by means of environmental parameters, such as phosphorus and nitrogen, BOD and OD concentrations. In Spearman's correlation, chlorophyll-a obtained a higher correlation with total phosphorus and in Random Forest, through feature importance analysis of each variable, the highest score was for BOD, which supported the idea that this algorithm when allied to the traditional statistic can present correlations not so observed in the same. In short, the results presented were acceptable according to the performance analysis metrics MAE (5.1914), RMSE (3.8473) and Oobscore (0. 2288). The values presented could be even better if more data were available, however, monitoring in the region is still recent.

Keywords: Chlorophyll-a; Radom Forest; Water Quality.

4.1 INTRODUÇÃO

A degradação do meio ambiente interfere diretamente na qualidade da água, e esta como recurso essencial para a vida e desenvolvimento de atividades econômicas não pode estar sujeita a contaminação. Neste sentido, monitorar e avaliar a qualidade da água são essenciais para o melhoramento das condições da mesma e o atendimento dos requisitos das legislações e dos atores interessados em uma bacia hidrográfica. No entanto, o entendimento das relações das variáveis ambientais e a origem de contaminações que alteram a dinâmica natural do corpo hídrico, não é algo trivial. Portanto, para que se tenha um melhor entendimento dos fenômenos sejam autóctones e alóctones em um corpo hídrico, utilizar-se de técnicas estatísticas, e outras mais modernas como aprendizado de máquinas (*machine learning*), são fundamentais para auxiliar na tomada de decisões (PARK *et al*., 2015; WANG *et al.*, 2017; CHOU *et al.*, 2018).

No contexto das tomadas de decisão, surgem as ferramentas computacionais e estatísticas que estão presentes em diversas áreas do conhecimento humano, e nas questões ambientais não é diferente. As correlações estatísticas permitem verificar relações entre os dados que por sua vez permitem inferir os tipos de contaminação que o corpo hídrico está exposto e quais as atividades são diretamente responsáveis. Sendo assim, suas aplicações facilitam esse entendimento. No entanto, muitas vezes o monitoramento da qualidade da água possui deficiências e limitações quanto aos registros e ao custo, o que acaba dificultando a aplicação e construção de modelos eficientes e que agreguem, de fato, o estudo dos corpos hídricos.

Dentre as diversas técnicas de aprendizado de máquinas, a *Random Forest* apresenta grande aplicabilidade e é de fácil utilização, sem a necessidade de normalização de dados, e ainda, consegue por meio da geração de árvores de decisão aleatórias minimizar o sobreajuste (*overfitting*). Além disso, a técnica permite a observação das variáveis mais importantes, de maior peso, na predição de uma variável, em um estudo ambiental, o que leva a modelos mais precisos e de representação facilitada para o público em geral.

Utilizando a técnica de Random Forest, proposta por Breiman (2001), Yajima e Derot (2018) propuseram o estudo da previsão de clorofila-a em um reservatório e um lago do Japão para

que fosse possível reduzir custos em análises desses corpos hídricos. Os autores observaram que há a necessidade de uma quantidade de dados relativamente maior para que as previsões de clorofila-a apresentem erros menores. Além disso, eles concluíram que a técnica apresentou, na análise de importância das variáveis relações fenomenológicas entre as variáveis utilizadas para a previsão de clorofila-a não apresentadas em outras técnicas, como por exemplo, na análise de correlação.

O uso de técnicas para previsão e predição de variáveis de qualidade água, de condições climáticas e ambientais vêm sendo aplicadas e obtendo bons resultados, gerando economia nos gastos de tratamento de água, no tratamento de doenças de origem ambiental e no custo de coletas em corpos hídricos, lênticos ou lóticos. Exemplo disso é a sua aplicação na predição de concentrações do pigmento clorofila-a, presente nos corpos hídricos, e que sua alta concentração é um dos indicadores de excesso de nutrientes, a eutrofização.

Neste sentido, identificar variáveis ambientais que mais influenciam o fenômeno de eutrofização é de extrema importância e uma boa ferramenta para os órgãos decisão direcionarem as medidas cabíveis. Diversas variáveis podem contribuir no floramento algal e na sua permanência em um corpo hídrico, especialmente em reservatórios. Sendo assim, estudar as variáveis e suas relações, utilizando de estatística e de algoritmos de aprendizado de máquinas, é uma alternativa para contribuir na mitigação de eventos indesejados como o aumento da trofia dos corpos hídricos (HOLLISTER *et al*., 2016; LI *et al*., 2017; KOVALENKO *et al*., 2018).

Sendo assim, este trabalho busca predizer a concentração de clorofila-a, em corpos hídricos do estado de Sergipe, utilizando a estatística e a técnica de aprendizado de máquinas *Random Forest* através de variáveis ambientais mensuradas em campo e em laboratório.

4.2 METODOLOGIA

O presente trabalho utilizou a linguagem de programação *Python*, com a utilização do *Software Jupyter* para utilização do modelo *Random Forest*. Trezentas árvores de decisão foi a quantidade utilizada nesse trabalho para a predição da clorofila-a por meio do algoritmo *Random Forest.*

Neste trabalho foi simulada a predição da clorofila-a de acordo com parâmetros ambientais: pH, Condutividade, Turbidez, Oxigênio dissolvido, Alcalinidade, Dureza, DBO_5, STD, Temperatura da Água, Cor, Fósforo total, Nitrato, Nitrito, Amônia, fósforo, coliformes e própria Clorofila-a. A diminuição do número de parâmetros de entrada do modelo foi utilizada para verificar a melhoria do mesmo e assim, reduzir o custo das análises de alguns desses parâmetros, bem como obter valores de concentração de clorofila-a sem a necessidade de uma análise laboratorial frequente deste parâmetro que, apesar de ser custosa, é essencial na avaliação da qualidade da água de corpos hídricos, principalmente, em reservatórios.

4.2.1 Área de Estudo e Seleção de dados

No estado de Sergipe, foram selecionados 18 reservatórios que abrangem quase todas as regiões. Na seleção de dados apenas um desses reservatórios teve que ser retirado devido à ausência de quantidade suficiente de parâmetros para a modelagem, que foi a barragem Amargosa que apresentou períodos de seca, impossibilitando a coleta de amostras. Os dados foram cedidos pelo Instituto Tecnológico e de Pesquisas do Estado de Sergipe e os reservatórios analisados estão apresentados na Tabela 7, bem como sua respectiva localidade. Esses reservatórios e riachos são utilizados para abastecimento, dessedentação animal, irrigação e recreação.

Tabela 7- Reservatórios avaliados e suas localidades

Reservatório	Cidade	Bacia
Algodoeiro	Nossa Senhora da Glória	Rio Vaza Barris
Lagoa do Rancho	Porto da Folha	Porto da Folha
Três Barras	Graccho Cardoso	Rio São Francisco
Comporta	Propriá	Rio São Francisco
Cumbe	Cumbe	Rio Japaratuba
Pau de Cedro	Nossa Senhora da Glória	Rio Vaza Barris
Coqueiro	Ribeirópolis	Rio Jacarecica
Macela	Itabaiana	Rio Sergipe
Jacarecica I	Itabaiana	Rio Jacarecica
Jacarecica II	Malhador	Rio Jacarecica
Poxim	São Cristóvão	Rio Poxim
Ribeira	Campo do Brito	Rio Traíras
Grutão de Carira	Carira	Rio Sergipe
Coité	Frei Paulo	Rio Vaza Barris
Taboca	Simão Dias	Rio Vaza Barris
Donísio Machado	Lagarto	Rio Piauí
Jabiberi	Tobias Barreto	Rio Jabiberi

A Figura 15 apresenta o mapa com as localizações dos corpos hídricos nas bacias hidrográficas do estado de Sergipe. O mapa foi elaborado utilizando a base de dados Geoespacial de Sergipe, elaborado com o apoio da Secretaria de Estado do Meio Ambiente e Recursos Hídricos (SEMARH/SE), utilizando o software livre QGis na versão 2.18.22.

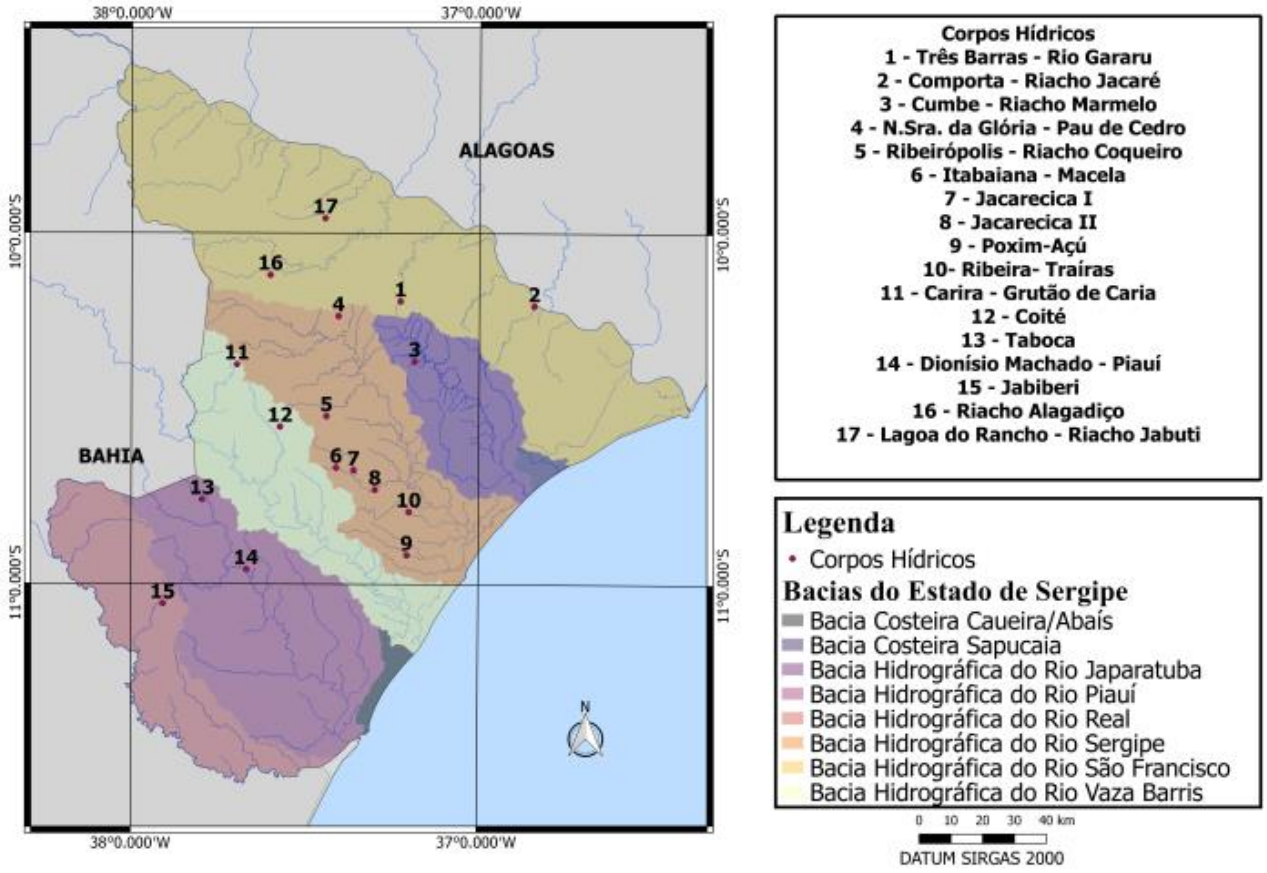


Figura 17-Mapa dos Corpo Hídricos de Sergipe

Para a construção do modelo, alguns filtros tiveram que ser utilizados. Dentre eles a eliminação de células vazias, que não possuíam medidas dos parâmetros, totalizando 137 linhas na base de dados. Essa seleção foi crucial para a seleção de quais variáveis entrariam no modelo que seriam as que mais tivessem, juntamente com a análise de clorofila-a. As variáveis selecionadas foram pH, condutividade, turbidez, oxigênio dissolvido, alcalinidade, dureza, DBO, Sólidos Totais Dissolvidos (STD), temperatura da água, cor aparente, fósforo total, nitrato, nitrito, amônia, fosfato, coliformes, clorofila-a, nitrogênio total. O modelo utilizou as 18 variáveis citadas

para a predição da clorofila-a para esses reservatórios. O monitoramento registrado pelo ITPS ocorreu entre os anos de 2013 e 2018.

4.2.2 Avaliação de Métricas

Com o objetivo de avaliar modelos propostos em sua capacidade de predição utilizou-se neste trabalho o cálculo de dois tipos de análise de performance: RMSE e MAE. O primeiro é o cálculo da raiz valor médio quadrado (RMSE) entre o valor real e o predito pelo modelo, e o segundo é o valor médio do erro absoluto (MAE) (NODOUSHAN, 2018; MONTEIRO e COSTA, 2018). Essas métricas de performance são as mais utilizadas para avaliar a acurácia de um modelo, quanto mais próximos de zero os valores de RMSE e MAE maior a capacidade de predição do modelo. As Equações 10 e 11 apresentam o cálculo do MAE e do RMSE, respectivamente, sendo que n é o número de dados referente a variável de saída, o target.

$$MAE = \frac{\sum_{i=1}^{n} \left| y_{i(observado)} - y_{i(predito)} \right|}{n} \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{i(observado)} - y_{i(predito)})^2}{n}} \quad (11)$$

Sendo: $y_{i(observado)}$=valor da variável mensurado, $y_{i(predito)}$=valor da variável predito pelo modelo, n=número de amostras.

Os valores de RMSE enfatizam a penalização de valores altos no modelo, enquanto o MAE apresenta os mesmos pesos para todos os erros e refletem a média dos desvios das predições. Além desses critérios, utilizou o ranqueamento de importância das variáveis característico do *Random Forest* para avaliar quais as variáveis que mais contribuíam para a predição da clorofila-a e para a, consequente, diminuição dos erros do modelo (LI *et al.*, 2018; CHAI e DRAXLER, 2014; CONVERTINO *et al.*, 2014).

4.3 RESULTADOS E DISCUSSÃO

4.3.1 Estatística Descritiva

As estatísticas descritivas dos corpos hídricos analisados estão apresentadas nas Tabelas 8 e 9. Os reservatórios apresentam uma faixa de pH entre 6 e 10,48, no entanto, 75% dos dados encontram-se em valores abaixo de 8,38. Valores de pH acima de 10 comprometem a vida aquática e causa a mortandade de peixes, além de contribuir na precipitação de metais. Essas máximas foram identificadas nos reservatórios Macela, Três Barras, Jacarecica I, Cumbe e Ribeira, todas no ano de 2018, no período seco no mês de fevereiro. Esses valores estão associados a floração de algas, devido a eutrofização como é o caso do reservatório Macela. Algas presentes em corpos hídricos, ainda, ocasionam alteração na alcalinidade do meio.

Tabela 8– Estatística descritiva dos dados dos reservatórios

	pH	**Condutividade (µS/cm)**	**Turbidez (uT)**	**Oxigênio dissolvido (mg/LO_2)**	**Alcalinidade (mg/L$CaCO_3$)**	**Dureza (mg/L$CaCO_3$)**	**DBO5 (mg/L O_2)**	**Clorofila a (µg/L)**
Dados	137	137	137	137	137	137	137	137
média	8,01	3.655,39	18,52	6,12	165,91	693,88	14,702	528,82
desvio	0,68	7.657,94	76,67	2,14	102,83	1.033,44	9,051	5.893,10
mín	6,12	27,65	0,11	0,00	1,69	4,02	0,000	0,00
25%	7,58	318,6	4,93	5,04	90,99	83,10	8,700	1,490
50%	7,98	1.233	8,50	6,28	160,90	227,90	13,700	10,10
75%	8,38	4.175	15,66	7,28	217,90	969,30	19,700	26,50
máx	10,48	70.930	899,40	11,77	621,90	6.106,00	44,68	69.000,00

Tabela 9– Estatística descritiva dos dados dos reservatórios

	STD (mg/L)	**Temperatura da Água (ºC)**	**Cor (uH)**	**Fósforo total (mg/L)**	**N-NO_3- (mg/L)**	**N-NO_2- (mg/L)**	**NH_3 (mg/L)**	**P-PO_4^{3-} (mg/L P)**	**Coliformes (NMP/100mL)**
Dados	137	137	137	137	137	137	137	137	137,00
Media	2.730,79	26,925	20,764	0,338	0,471	0,105	0,351	0,198	2,35E+10
Desvio	5.593,12	1,818	137,804	0,556	0,755	0,460	1,300	0,824	1,59E+11
Mín	81,00	22,000	0,020	0,010	0,000	0,000	0,020	0,010	1,80E+06
25%	260,40	22,500	1,090	0,040	0,070	0,010	0,050	0,010	2,00E+07
50%	967,00	27,000	2,070	0,070	0,180	0,010	0,100	0,020	1,10E+08
75%	3.030,14	28,000	5,490	0,500	0,480	0,030	0,180	0,030	5,40E+08
máx	48.940,00	32,500	1.602,000	3,210	4,230	4,450	11,940	8,460	1,60E+12

A condutividade dos corpos hídricos registrou seu maior no açude Três Barras. Nesse açude, há o uso para psicultura que contribui para a presença de sal na água, já que uma quantidade mínima é necessária para manter o balanço osmótico dos peixes, conforme Stone *et al.* (2013). Apesar dos maiores valores em condutividade, o açude Três Barras não apresentou os maiores em Sólidos Totais Dissolvidos. Os maiores valores de STD, foram encontrados na Lagoa do Rancho, que neste mesmo período obteve valores muito elevados de condutividade e dureza, no ano de 2013, no fim do período seco. Além disso, neste ponto os valores de clorofila-a chegaram a mais de 50 µg/L, apresentando assim evidências de contaminação, acrescido de níveis baixos de oxigênio dissolvido, beirando o estado de hipoxia (< 3 mg/L).

A clorofila-a, variável a ser predita nesse estudo, obteve em 75% dos dados valores próximos de 26,5 µg/L, o que para a resolução do CONAMA 357/2005, não enquadraria os corpos hídricos com classe 1, já que o limite é de 10 µg/L, mas na classe 2, em que há necessidade de tratamento de água convencional para que seja possível o consumo humano. No entanto, o limite para a classe 2 é de 30 µg/L, o que deixa a grande maioria das amostras analisadas próxima a classe 3, tendo a necessidade de um tratamento avançado. O limite para a classe 3 é de 60 µg/L, no entanto, vários riachos e reservatórios obtiveram valores bem acima, como o caso do reservatório da Macela, lagoa do Rancho, Riacho Coqueiro, Cumbe, Pau de Cedro e Algodoeiro. Os maiores valores foram registrados no reservatório Cumbe, que na maioria dos registros possuiu valores acima de 100 µg/L, que é utilizado para a agricultura e acaba sendo impactado com grande quantidade de nutrientes, bem como o Reservatório Pau de Cedro, em Nossa Senhora de Glória, que é utilizado para abastecer o perímetro irrigado da região. Esse reservatório pode ser classificado como classe IV, devido a sua degradação, conforme ressalta Melo *et al.* (2015).

As concentrações dos nutrientes analisados NO_2, NO_3, NH_4, P e PO_4, são fatores que contribuem para a eutrofização do corpo hídrico e ainda acarretam diminuição dos níveis de oxigênio em seus processos oxidativos, causando grandes prejuízos aos corpos hídricos interferindo na dinâmica natural dos mesmos. Em ambientes lênticos, devido, principalmente, ao tempo de residência, há uma maior ocorrência do fenômeno de eutrofização. As maiores concentrações de fosfato foram encontradas no riacho da Macela em que houve o maior valor 8,46 mg/L no final do período seco no ano de 2013, e em 2018 a concentração foi 1,17 mg/L. Esses

valores indicam que este corpo hídrico sofre com efeito do mau uso de fertilizantes e pesticidas que possuem fósforo em sua essência. Os reservatórios de Cumbe, Pau de Cedro, Ribeira e a comporta do Riacho Jacaré, também, obtiveram valores alto desse parâmetro indicando impactos advindos de esgoto doméstico e atividade agrícola na região.

Fósforo total foi observado com valores muito acima da resolução CONAMA 357/2005 de 0,03 mg/L em ambientes lênticos, classe 2. Os reservatórios que mais se destacam são o reservatório Macela, Cumbe, Pau de Cedro e a comporta do riacho Jacaré. Esses reservatórios sofrem com contaminação de esgoto, lixiviação dos solos que contribuem para o enriquecimento de nutrientes no ambiente interferindo no uso para abastecimento e até recreação, já que há aparecimento de algas e macrófitas. Para as concentrações dos compostos nitrogenados, o reservatório de Carira, o reservatório Coité e Lagoa do Rancho, possuem valores de 1 mg/L a 4,23 mg/L indicando um valor muito alto, principalmente, se observamos que estamos nos referindo apenas a uma forma do nitrogênio em todo o seu ciclo. O açude Coité ainda apresentou valores de amônia elevados, 3,73 mg/L, no mesmo período que houve seu pico em nitrato, no final do mês de junho de 2013, período chuvoso da região, possivelmente, o que ocasionou lixiviação das terras cultivadas.

As concentrações de Demanda Bioquímica de Oxigênio (DBO), segundo Libânio (2010), é a quantidade de oxigênio consumida para a estabilização da matéria orgânica por meio de bactérias. Os valores elevados de DBO indicam contaminação do corpo hídrico, e, portanto, indícios de redução de Oxigênio Dissolvido (OD). Para corpos hídricos classe I não pode ser maior que 3 mg/L, classe II até 5mg/L e classe 3 até 10 mg/L. Observa-se que 75% das amostras possuem valor abaixo de 20mg/L, indicando contaminação, principalmente de esgotos, e corpos hídricos classe 3 e 4. O riacho Jacaré, obteve um registro de Oxigênio dissolvido próximo a 1 mg/L, e DBO igual a 26,3 mg/L de O_2. A comporta do Riacho Jacaré, ainda, chegou a obter níveis de anoxia, em que não houve registro de OD no meio. Este riacho sofre com deposição de esgoto, mas principalmente contaminação de fertilizantes, o que acaba ocasionando déficit de oxigênio devido aos processos oxidativos como a nitrificação, conforme observado por Macedo *et al.* (2010) e Lucas *et al.* (2014).

A temperatura da água é fundamental para as reações que nela ocorre e para a manutenção da vida. Os valores de analisados estão entre 22 e 32 ºC, valores normais para os corpos hídricos do estado de Sergipe. Os menores valores foram observados no período chuvoso e os maiores no período seco. Em relação aos coliformes fecais, os altos valores estão relacionados aos esgotos, drenagem urbana e fezes animais que acabam interferindo na dinâmica natural dos corpos hídricos. Essa contaminação se intensifica nos períodos chuvosos em que ocorrem escoamento superficial dos centros urbanos para os corpos hídricos aumentando a quantidade de coliformes no meio, bem como advindo de estações de tratamento de esgoto próximas e de esgotamento sem tratamento prévio, conforme ressaltam Buzelli e Cunha-Santino (2013), Mohammed *et al.* (2017) e Avila *et al.* (2018). A resolução do CONAMA 357/2005 limita a 200 NMP/100mL para classe 1, 1000NMP/100mL para classe 2, 4000 NMP/100mL para classe 3 e valores acima deste último pertencentes a classe 4. Observa-se contaminação por coliformes em todos os corpos hídricos analisados. Os reservatórios da Macela, Algodoeiro, Três Barras, Jacarecica I e a comporta do riacho Jacaré possuem os valores mais elevados de coliformes chegando a 920000 NMP/100mL mostrando contaminação elevada de tais corpos hídricos por esgoto, além de alguns desses sofrerem com a eutrofização.

Cabe ainda ressaltar que corpos hídricos lênticos sofrem com o processo inexorável de assoreamento, conforme Libânio (2010), explicando os elevados valores de turbidez e sólidos totais dissolvidos (STD). A elevação desses dois últimos parâmetros citados, portanto, não se deve, apenas as atividades antrópicas que agridem esse corpo hídrico.

4.3.2 Correlações entre as variáveis e clorofila-a

Para avaliação da correlação das variáveis com a clorofila-a, primeiramente, fez-se uso da correlação de Spearman. Neste método, quanto maior o valor da variável, mais próximo de +1, em relação a outra, maior sua correlação, e mais próxima de -1, menor sua correlação. Observa-se que as variáveis de maior correlação com a clorofila-a foram o fósforo total e sólidos totais dissolvidos. Isso representa um aumento gradativo das quantidades de fósforo total devido as atividades desenvolvidas nas margens desses reservatórios e riachos. Atividades como a agricultura é uma geradora do fenômeno de eutrofização, pois contém nutrientes para as algas. Além disso, o esgoto descartado *in natura* colabora de forma direta colaboram para o aumento de fósforo no corpo

hídrico. O valor estipulado para o fósforo pela resolução 357/2005 do CONAMA é de 0,03 mg/L em ambientes lênticos e esse valor nos corpos hídricos analisados tem um máximo que ultrapassa em 100 vezes esse valor, no Riacho Jacaré, no ano de 2017, final do período chuvoso.

A segunda maior correlação foi com os sólidos totais dissolvidos (STD). Essa correlação deve-se, principalmente, a presença de algas em corpos eutrofizados que contabilizam como sólidos suspensos que é uma parte do STD, similar ao encontrado no trabalho de Li *et al.* (2017). Além disso, a presença de sólidos afeta a fotossíntese, pois, impede a penetração de luz no corpo hídrico.

Alcalinidade, cor e amônia, obtiveram uma alta correlação. A alcalinidade varia muito com o aumento do crescimento algal, já que CO_2 é gerado durante a fotossíntese dos organismos fotossintéticos. Além disso, a presença de algas altera a cor, odor e gosto dos corpos hídricos. A amônia é um componente nitrogenado que no ciclo nitrogenado é indicador de contaminação recente nos corpos hídricos

A temperatura teve uma relação inversa e menor com a clorofila-a conforme observado na Figura 16. As variações de temperatura em regiões tropicais e áridas não produzem grande efeito no fenômeno de eutrofização, já que são muito pequenas. Outra variável com baixa correlação foi o oxigênio dissolvido, já que a uma relação inversa, menor a quantidade de oxigênio, tende-se observar maior crescimento algal. Além disso, as variações da concentração de oxigênio observadas não foram muito elevadas.

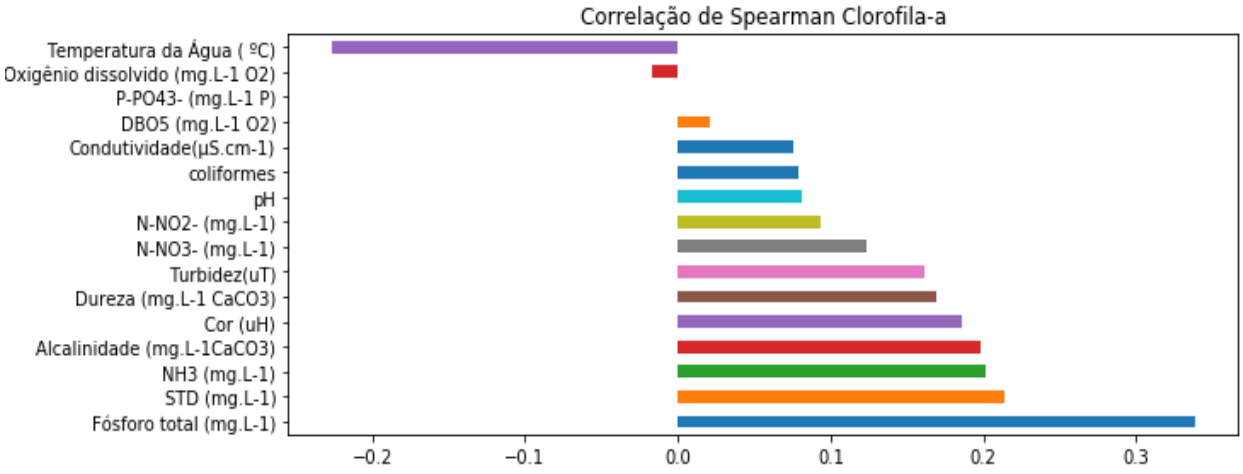


Figura 18 - Correlação Spearman da clorofila-a com as demais variáveis analisadas

4.3.3 Random Forest

Um dos grandes potenciais dos algoritmos de *machine learning* é facilitar o entendimento dos fenômenos e identificar relações entre as variáveis não observadas ou que não estavam claras no uso da estatística convencional, conforme argumentam Yajima e Derot (2017). Nesse sentido, algumas relações não identificadas na matriz de correlação de Spearman podem ocorrer no ranking de importância das variáveis (*feature importance*) para a predição de clorofila-a. A importância das variáveis no método *Random Forest* é não linear e a de Spearman é monotônica, ou seja uma função que preserva as suas relações. Se ela crescente se mantém crescente, assim como, se for decrescente a mesma se mantém decrescente sejam estas lineares ou não.

Neste trabalho, em uma primeira simulação, Figura 17, foram utilizadas todas as variáveis analisadas, além, da variável categórica local. Nessa análise foi realizada o ranqueamento de *feature importance* para a predição da clorofila-a, a variável Nitrogênio Kjedahl, também fora adicionada, como Nitrogênio Total, no entanto, não obteve um rankeamento muito alto. Observou-se como na correlação de Spearman um maior valor atribuído ao fósforo total, como variável mais importante para a predição da clorofila-a. Além disso, a DBO já obteve um maior peso em relação a correlação analisada no tópico anterior.

As variáveis pH e dureza obtiveram também pesos consideráveis na predição da clorofila. O pH tende a variar em corpos hídricos eutrofizados devido ao aumento do CO_2 na água, produto este advindo da fotossíntese das plantas. O aumento das concentrações de CO_2 acaba trazendo um aumento da dureza da água devido aos carbonatos formados. Além dessas variáveis, nitrito e nitrato, advindos principalmente, de fertilizantes que com o assoreamento e a lixiviação do solo, que carregam os sólidos (estes também bem ranqueados), acabam escoando para o corpo hídrico.

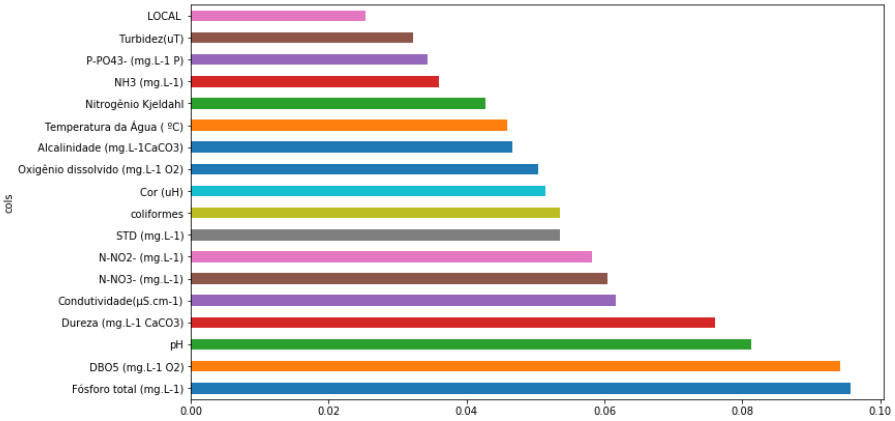


Figura 19- *Feature importance* com variável categórica

Em uma segunda tentativa, Figura 18, foram retiradas variáveis de menor peso como Local, Turbidez, fosfato, amônia e o nitrogênio Kjedahl. Nesse resultado, a DBO passou a observar um maior peso em relação a predição de clorofila-a. A DBO relaciona-se a despejos orgânicos em um corpo hídrico, característico de ambientes eutrofizados, conforme comentam Li *et al.* 2017(b) e Jardim (2011). Os corpos hídricos que mais possuem essa relação de forma preponderante são a Lagoa do Rancho, o Reservatório Ribeira, Coité, Comporta, Jacarecica II e Macela. Esse último, já citado em diversos trabalhos como um dos corpos hídricos mais eutrofizados no estado de Sergipe conforme Sena *et al.* (2012), Garcia *et al.* (2017) e Santos *et al.* (2017).

Algumas variáveis mudaram a ordem como o pH e o Fósforo total, além dos coliformes que passou a ter um maior peso. Coliformes referem-se à contaminação orgânica e está relacionada a DBO do corpo hídrico. As demais variáveis praticamente, mantiveram a ordem em relação a primeira simulação, e pesos bem próximos uns dos outros.

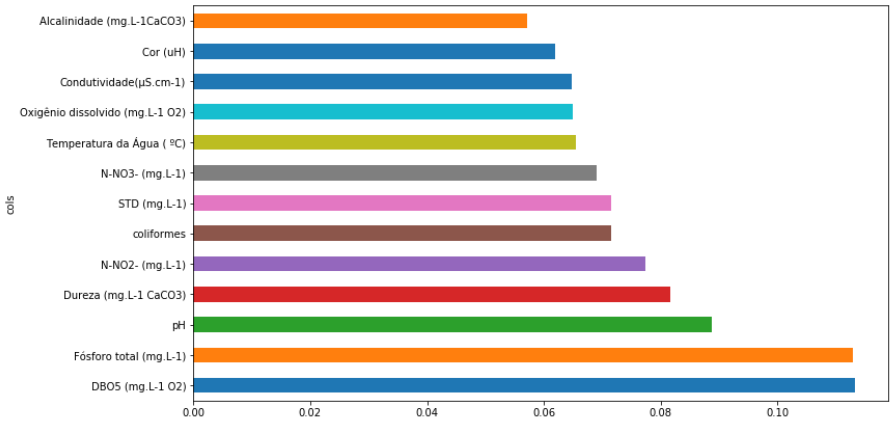


Figura 20- Segunda tentativa *feature importance*

Conforme afirmado por Hollister *et al.*, (2016) e Yajima e Derot, (2017) o uso da técnica de *machine learning* possibilita observar relações com a variável *target* que não se sobressaíram na correlação de Spearman. A DBO que não estava entre as variáveis com maior correlação com a clorofila-a observou o valor próximo ao do fósforo total como as duas variáveis de maior importância para a predição de clorofila-a nos corpos hídricos analisados. O pH, também, destaca-se como a terceira mais importante para a predição dos valores de clorofila-a. Esse resultado indica que a combinação dos dois métodos favorece o entendimento e um direcionamento mais preciso na tomada de decisão para intervenções nesses corpos hídricos visando o melhoramento da qualidade da água dos mesmos. Por fim, esses resultados permitem ainda identificar como as atividades antrópicas e fenômenos naturais impactam diretamente nos recursos hídricos, aumentando a responsabilidade de cada um em suas ações.

4.3.4 Análise de Métricas

A Tabela 10 apresenta as métricas RMSE e MAE para os dois modelos de *Random Forest* aplicados neste trabalho. A primeira simulação refere-se ao uso de todas variáveis incluindo Nitrogênio Kjedahl e o Local. A segunda simulação refere-se a retirada das variáveis de menor peso como Local, Turbidez, fosfato, amônia e o nitrogênio Kjedahl. A terceira simulação corresponde a retirada de todas as variáveis identificadas nos dois primeiros e a alcalinidade. No entanto, conforme resultados dos erros MAE e RMSE, observou-se que não houve uma melhoria considerável do algoritmo ao realizar esse procedimento. Dessa forma os erros para o conjunto de dados avaliados apresentam erros dentro do aceitável. Li *et al.* (2018), ao predizer clorofila-a no

Lago Baiyangdian encontraram valores de MAE entre 7 e 11, variando as variáveis de entrada do algoritmo de *machine learning Random Forest*. Li *et al.* (2017) observaram valores entre 1,74 e 3,27 para a clorofila-a no lago Poyang na China. Os autores utilizaram uma base de dados muito maior que a deste trabalho e, neste sentido, os resultados são considerados aceitáveis.

Tabela 10- Análise das Métricas do *Random Forest* para a predição de Clorofila-a (μg/L)

Simulações	***MAE Treino***	***RMSE Treino***	***MAE Teste***	***RMSE Teste***	***OOB Score***
1	6,0989	4,4102	5,5461	4,1751	0,1945
2	6,0595	4,3139	5,1914	3,8473	0,2288
3	6,0447	4,3062	5,1499	3,9164	0,2347

Os valores do OOBscore, na Tabela 10, referem-se ao valor das linhas não incluídas no treinamento comparadas ao predito pelo algoritmo *Random Forest*, fazendo dessa forma uma validação cruzada, com 1/3 dos dados utilizados. Os valores oscilaram pouco mostrando precisão do modelo igualmente ao obtido por Li *et al.* (2018). Apesar de um aumento do valor do OOBscore quando a quantidade de variáveis foi reduzida, não se observou uma alteração significativa nos valores de MAE e RMSE, o que indica que manter o segundo modelo é o mais adequado para a aplicação neste trabalho.

A análise de predição com 31 dados aleatórios da base de dados para o valor da clorofila-a, conforme apresentado na Tabela 11. Os valores foram comparados e apresentaram similaridade na média e no desvio padrão. A grande aproximação em 75% dos dados (terceiro quartil) com valores de concentração de clorofila-a abaixo de 21 μg/L, reforçam a precisão e similaridade do algoritmo *Random Forest* com os dados de campo. A Figura 19 apresenta uma percepção gráfica da proximidade dos valores reais e preditos modelo em que se observa uma proximidade significativa entre eles.

Tabela 11- Comparativo estatístico entre as concentrações de clorofila-a

	Valores Reais (µg/L)	*Valores Preditos (µg/L)*
dados	31	31
média	14,7464	14,0916
desvio padrão	16,9587	12,4235
mín	0,0000	0,8223
25%	1,9950	4,8934
50%	7,9000	8,5855
75%	21,2500	20,7021
máx	57,1000	44,2514

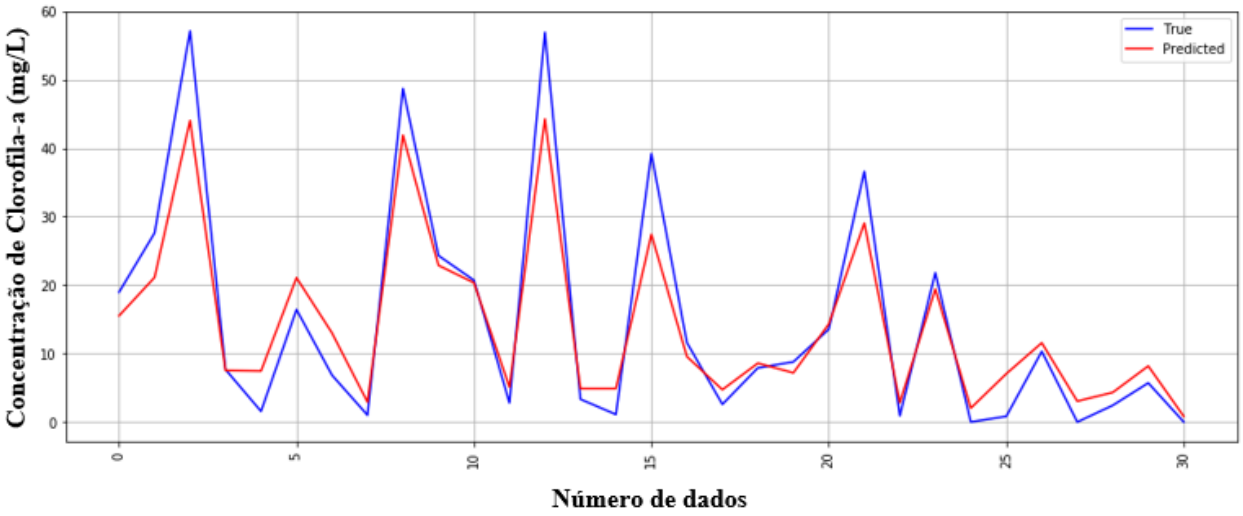


Figura 21 - Valores reais e preditos de concentração de clorofila-a

4.3.5 Inferências

Esse trabalho ressalta que as coletas foram espaçadas, necessitando uma maior regularidade das mesmas o que poderia melhorar os resultados e permitir análises melhores dos impactos sazonais nas amostras e na predição de clorofila-a. O desenvolvimento de programas de monitoramento necessita dessa regularidade para identificação de anomalias nos resultados, além, de um melhor direcionamento na tomada de decisões, uma das sugestões do artigo de Peletz *et al.* (2018). Um maior conjunto de dados permitiria análises direcionadas para cada corpo de forma

mais aprofundada e um algoritmo com erros menores, embora estes ainda estejam dentro do aceitável para os dados disponíveis.

Nesse sentido, um caminho que poderia ser adotado neste trabalho para aprimorar os resultados seria interpolar os valores para um aumento da quantidade de dados. No entanto, isso implicaria na criação de dados o que traria diversos desvios, principalmente, porque não havia periodização nas coletas. Além disso, existe a necessidade da melhoria do programa de monitoramento do estado de Sergipe, de forma a facilitar o entendimento e o uso dos dados por terceiros que busquem contribuir no entendimento e diagnostico dos corpos hídricos.

Inegável que a predição da variável clorofila-a, no contexto da qualidade de água, é fundamental para o entendimento da eutrofização dos corpos hídricos. A degradação da qualidade da água reflete em impactos como gastos no tratamento da água, perda de valor do corpo hídrico, além, de potencializar a possibilidade de doenças transmitidas pela água. Observando os efeitos indiretos da eutrofização, pode-se pensar no impacto das doenças que afastam o trabalhador da sua função, diminuindo a produtividade econômica. Portanto, mitigar os vetores da eutrofização é algo crucial para os gestores. Fatores como a decomposição da vegetação natural existente antes da implementação de reservatório, por exemplo, acabam refletindo na depleção de oxigênio do corpo hídrico, o que favorece a condição para a eutrofização conforme afirmam Libanio (2010) e Tundisi e Matusmara-Tundisi (2011), além de Hollister *et al.* (2017) e Avila *et al.* (2018).

4.4 CONCLUSÃO

Neste trabalho, foram avaliados corpos hídricos do estado de Sergipe, buscando prever a concentração de clorofila-a a partir de parâmetros de qualidade da água. Neste sentido, foram avaliadas as correlações estatísticas destas variáveis com a clorofila-a e a predição desta pela técnica de *machine learning, random forest.*

O modelo de *machine learning, random forest* contribuiu na ampliação do entendimento dos fenômenos do corpo hídrico, das variáveis essenciais na predição de clorofila-a e de como elas se relacionam. O algoritmo ainda apresenta como vantagem a não necessidade de transformação

dos dados, como normalização para a sua aplicação mostrando sua melhor aplicabilidade do modelo frente aos demais que necessitam desse passo antes de iniciar uma análise.

Dessa maneira, foram comparadas as variáveis de maior correlação com a clorofila-a estatisticamente e pela técnica *random forest*. Observou-se diferenças nas variáveis com maior peso para a predição nas duas técnicas. Na correlação de Spearman, o fósforo total obteve a maior correlação entre as variáveis analisadas, devido, principalmente, as suas altas concentrações em alguns corpos hídricos analisados, valores estes acima dos limites previstos da legislação. Na análise de *feature importance* do *random forest*, a DBO apresentou o maior peso para a predição de clorofila-a, e o fósforo total com o segundo maior peso. O maior peso para a DBO é observado devido a presença de matéria orgânica em corpos hídricos e dos descartes advindos de efluentes domésticos que corroboram para aumento das concentrações de clorofila-a, DBO e fósforo total.

Sendo assim, os baixos valores nos erros avaliados de predição, no que se refere aos dados deste trabalho, mostram a eficiência do modelo de predição da clorofila-a, com base nas variáveis que já são medidas no estado. A técnica, portanto, pode auxiliar a tomada de decisão, reduzir custos de campo e auxiliar o programa de monitoramento. Essa aplicação não se restringe aos corpos lênticos, mas pode ser usada para os corpos lóticos do estado. Uma base maior de dados pode aumentar ainda mais a capacidade de predição do algoritmo.

Este trabalho, buscou apresentar para as autoridades a necessidade de um programa de monitoramento mais efetivo, embora no estado este ainda seja recente, e a importância do mesmo para a sociedade. Ao realizar um programa de monitoramento é fundamental sua regularidade e menor espacialidade entre as medidas, com maior frequência de amostragem. Além disso, um registro mais cuidadoso dos dados deve ser feito para que usuários secundários possam analisá-los e obter informações que auxiliem a tomada de decisão, apresentando estes resultados aos órgãos responsáveis.

Sendo assim, pode-se concluir que técnicas de *machine learning* contribuem para análise dos corpos hídricos, na avaliação da qualidade da água e auxiliam na identificação de correlações entre as variáveis que muitas vezes são consideradas relativamente baixas em técnicas de correlação tradicionais, permitindo, assim, averiguar relações despercebidas ou eclipsadas pela

estatística tradicional. Em trabalhos futuros, sugere-se a incorporação de outras variáveis para a predição de clorofila-a como o Carbono Orgânico Total, bem como ampliação do conjunto de dados trabalhada visando um aperfeiçoamento do modelo ou utilização de outros modelos.

REFERÊNCIAS

AVILA, R.; HORN, B.; MORIARTY, E.; HODSON, R.; MOLTCHANOVA, E. Evaluating statistical model performance in water quality prediction. **Journal of environmental management**, *206*, 910-919, 2018.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BUZELLI, G. M; CUNHA-SANTINO, M. B.Análise e diagnóstico da qualidade da água e estado trófico do reservatório de Barra Bonita, SP. **Revista Ambiente & Água - An Interdisciplinary Journal of Applied Science**. v. 8, n.1, 2013.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. **Geoscientific model development**, v. 7, n. 3, p. 1247-1250, 2014.

CONVERTINO, M.; MUÑOZ-CARPENA, R.; CHU-AGOR, M. L.; KIKER, G. A.; LINKOV, I. Untangling drivers of species distributions: Global sensitivity and uncertainty analyses of MaxEnt. **Environmental Modelling & Software**, v. 51, p. 296-309, 2014.

CHOU, J.S.; HO, C.C.; HOANG, H.S. Determining quality of water in reservoir using machine learning. **Ecological informatics**, v. 44, p. 57-75, 2018.

GARCIA, C.A.B; GARCIA, H.L.; MENDONÇA, M.C.S.; SILVA, A.F.; ALVES, J.P.H.; COSTA, S.S.L.; ARAÚJO, R.G.O.; SILVA, I.S. Assessment of Water Quality Using Principal Component Analysis: A Case Study of the Açude da Macela, Sergipe, Brazil. **Modern Environmental Science and Engineering**, V.3, No. 10, pp. 690-700, 2017.

HOLLISTER, J. W.; MILSTEAD, W. B.; KREAKIE, B. J. Modeling lake trophic state: a *Random Forest* approach. **Ecosphere**, v. 7, n. 3, 2016.

JARDIM, B.F. M.. Variação dos parâmetros físicos e químicos das águas superficiais da bacia do Rio das Velhas-MG e sua associação com as florações de cianobactérias. - Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia, 2011.

KOVALENKO, K. E.; REAVIE, E. D.; BARBIERO, R. P.; BURLAKOVA, L. E.; KARATAYEV, A. Y.; RUDSTAM, L. G.; WATKINS, J. M. Patterns of long-term dynamics of aquatic communities and water quality parameters in the Great Lakes: Are they synchronized? **Journal of Great Lakes Research**, *44*(4), 660-669, 2018.

LI, B.; YANG, G.;WAN, R.; HÖRMANN, G.; HUANG, J.; FOHRER, N.; ZHANG, L. Combining multivariate statistical techniques and random forests model to assess and diagnose the trophic status of Poyang Lake in China. **Ecological Indicators**, *83*, 74-83, 2017.

LI, X.; SHA, J.; WANG, Z.L. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. **Environmental Science and Pollution Research**, p. 1-11, 2018.

LI, X.; SHA, J.; WANG, Z.L. Chlorophyll-a prediction of lakes with different water quality patterns in China based on hybrid neural networks. **Water**, v. 9, n. 7, p. 524, 2017. (b)

LIBÂNIO, M. **Fundamentos de qualidade e tratamento de água**. Átomo, 2010.

LUCAS, A.A.T.; MOURA, A.S.A; NETTO, A.O.; FACCIOLI, G.G.; SOUSA, I.F. Qualidade da água no riacho Jacaré, Sergipe, Brasil usada para irrigação. **Revista Brasileira de Agricultura Irrigada** v.8, nº2, p.98-105, 2014.

MACEDO, F.L; PEDRA, W.N.; MELLO JUNIOR, A.V. Caracterização Fisiográfica da sub-bacia do Riacho Jacaré-SE. **Revista Brasileira de Geografia Física**, ed 03, 2010.

MELO, A. P. S.; GARCIA, H.L.; MENDONÇA, M.C.S.; BARRETO, V.L.; GARCIA, C.A.B. Qualidade da água dos reservatórios Algodoeiro e Glória através do índice de qualidade de água de reservatório. **XXI Simpósio Brasileiro de Recursos Hídricos**, Brasília, Brasil, 2015.

MOHAMMED, H.; HAMEED, I. A.; SEIDU, R. *Random Forest* tree for predicting fecal indicator organisms in drinking water supply. In: **Behavioral, Economic, Socio-cultural Computing (BESC), 2017 International Conference on**. IEEE, 2017. p. 1-6.

MONTEIRO, M.; COSTA, M. A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. **Hydrology**, v. 5, n. 3, p. 1-20, 2018.

NODOUSHAN, E.J. Monthly Forecasting of Water Quality Parameters within Bayesian Networks: A Case Study of Honolulu, Pacific Ocean. **Civil Engineering Journal**, v. 4, n. 1, p. 188-199, 2018.

PARK, Y.; CHO, K. H.; PARK, J.; CHA, S. M.; KIM, J. H. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. **Science of the Total Environment**, *502*, 31-41, 2015.

PELETZ, R.; KISIANGANI, J.; BONHAM, M.; RONOH, P.; DELAIRE, C.; KUMPEL, E.; MARKS, S.; KHUSH, R. Why do water quality monitoring programs succeed or fail? A qualitative comparative analysis of regulated testing systems in sub-Saharan Africa**. International journal of hygiene and environmental health**, 2018

SANTOS, C.E.O; PEIXOTO, J.S.; ALVES, J.P.H. Geoquímica das águas do reservatório Poção da Ribeira, Agreste Central de Sergipe. **Scientia Plena**, v. 13, n. 10, 2017.

SENA, I. M. N.; MACEDO, L. C. B.; ALVES, J.P.H. Qualidade Da Água Do Reservatório Macela/Itabaiana-Sergipe 2004-201. 2º Congresso Internacional - Resag, 2015.

STONE, N. M.; SHELTON, J. L.; HAGGARD, B. E.; THOMFORDE, H. K. **Interpretation of water analysis reports for fish culture**. Southern Regional Aquaculture Center, 2013.

TUNDISI, José Galizia; MATSUMURA-TUNDISI, Takako. **Recursos hídricos no século XXI**. Oficina de Textos, 2011.

WANG, Xiaoping; ZHANG, Fei; DING, Jianli. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. **Scientific reports**, v. 7, n. 1, p. 12858, 2017.

YAJIMA, H.; DEROT, J. Application of the *Random Forest* model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. **Journal of Hydroinformatics**, v. 20, n. 1, p. 206-220, 2018.

5 CONCLUSÃO GERAL

O presente trabalho possibilitou uma abordagem atual para a análise da qualidade da água por meio de técnicas de aprendizagem de máquinas. O problema de eutrofização é recorrente em diversos lugares do mundo devido a atividades antrópicas que acabam causando danos que muitas vezes são irremediáveis aos corpos hídricos. Isso foi verificado nos estados de Sergipe e São Paulo, no Brasil, e em 4 estados dos EUA (Geórgia, Flórida, Alabama e Arizona). No primeiro artigo, observou-se que o uso das séries temporais nas Redes Neurais Artificiais apresentou resultados bem precisos e que é uma ferramenta exequível usada na predição de concentração de clorofila-a. No segundo artigo, referente aos reservatórios no estado de Sergipe, a adaptação do algoritmo *random forest* a uma base de dados menor propiciou boas correlações para a variável clorofila-a e precisão dos resultados.

Além disso, os resultados mostraram a importância de se manter programas de monitoramento com base de dados organizadas para consulta e, principalmente, para facilitar a análise o que tornaria a tomada de decisão mais efetiva e rápida. Vale ainda reforçar que os algoritmos são ferramentas de auxílio e não tomadores de decisão por si só e estes mostram os caminhos a serem traçados com maior facilidade, mas a decisão e a atitude advêm da inciativa humana.

Por fim, um trabalho como esse vem alertar a necessidade de educação ambiental para a população, da consciência de todos os setores da sociedade que utilizam os recursos hídricos sem pensar se está prejudicando ou não, e como uma ferramenta computacional pode auxiliar nas alternativas a serem seguidas no sentido de manter a qualidade da água para que este recurso que pareceu um dia ser infinito, não se torne cada dia mais finito.